

텍스트 마이닝





힙합가사 텍스트 마이닝








이해하기

- ❖ 문자로 된 데이터에서 가치 있는 정보를 얻어 내는 분석 기법
- ❖ SNS나 웹 사이트에 올라온 글을 분석해 사람들이 어떤 이야기를 나누고 있는지 파악할 때 활용
- ❖ 형태소 분석(Morphology Analysis) : 문장을 구성하는 어절들이 어떤 품사로 되어 있는지 분석
- ❖ 분석 절차
 - 형태소 분석
 - 명사, 동사 형용사 등 의미를 지닌 품사 단어 추출
 - 빈도표 만들기
 - 시각화

Java 설치

❖ <https://www.java.com/ko/download/manual.jsp>

 Windows  무엇을 선택해야 할까요?

 <u>Windows 온라인</u> 파일 크기: 1.99 MB	<u>지침</u>	Java를 설치한 후 브라우저에서 Java를 사용으로 설정하려면 브라우저를 재시작해야 할 수 있습니다.
 <u>Windows 오프라인</u> 파일 크기: 69.61 MB	<u>지침</u>	
 <u>Windows 오프라인 (64비트)</u> 파일 크기: 79.19 MB	<u>지침</u>	

32비트 및 64비트 브라우저를 교대로 사용하는 경우, 각 브라우저에 대해 Java Plug-in이 필요하므로 32비트 Java와 64비트 Java를 모두 설치해야 합니다. » [Windows용 64비트 Java에 대한 FAQ](#)

패키지 설치 및 로드

패키지 설치

```
install.packages("rJava")
```

```
install.packages("memoise")
```

```
install.packages("KoNLP")
```

```
install.packages("remotes")
```

```
remotes::install_github('haven-jeon/KoNLP', upgrade="never",  
INSTALL_opts=c("--no-multiarch"))
```

패키지 로드

```
library(KoNLP)
```

```
library(dplyr)
```

패키지 로드 에러 발생할 경우 - java 설치 경로 확인 후 경로 설정

java 폴더 경로 설정

```
Sys.setenv(JAVA_HOME="C:/Program Files/Java/jre1.8.0_111/")
```

특수문자 제거

```
# 데이터 불러오기
```

```
txt <- readLines("hiphop.txt", encoding = "UTF-8")
```

```
head(txt)
```

```
#문자열 처리 관련 라이브러리 설치
```

```
install.packages("stringr")
```

```
library(stringr)
```

```
# 특수문자 제거
```

```
txt <- str_replace_all(txt, "WWW", " ")
```

명사 추출

```
# 명사 추출하기
extractNoun("대한민국의 영토는 한반도와 그 부속도서로 한다")
# 가사에서 명사추출
nouns <- extractNoun(txt)
# 추출한 명사 list를 문자열 벡터로 변환, 단어별 빈도표 생성
wordcount <- table(unlist(nouns))

# 데이터 프레임으로 변환
#stringsAsFactors = F 이므로 문자열 처리
df_word <- as.data.frame(wordcount, stringsAsFactors = F)
#확인
df_word
```

데이터 가공

```
# 변수명 수정
df_word <- rename(df_word,
                  word = Var1,
                  freq = Freq)
#filter, %>% 사용 위해 등록
library(dplyr)
# 두 글자 이상 단어 추출
df_word <- filter(df_word, nchar(word) >= 2)
#상위 20개 추출
top_20 <- df_word %>%
  arrange(desc(freq)) %>%
  head(20)
#확인
top_20
```


워드 클라우드

패키지 설치

```
install.packages("wordcloud")
```

패키지 로드

```
library(wordcloud)
```

```
library(RColorBrewer)
```

워드 클라우드

```
# Dark2 색상 목록에서 8개 색상 추출  
pal <- brewer.pal(8,"Dark2")
```

```
set.seed(1234)                # 난수 고정  
wordcloud(words = df_word$word, # 단어  
           freq = df_word$freq, # 빈도  
           min.freq = 2,        # 최소 단어 빈도  
           max.words = 200,     # 표현 단어 수  
           random.order = F,    # 고빈도 단어 중앙 배치  
           rot.per = .1,       # 회전 단어 비율  
           scale = c(4, 0.3),  # 단어 크기 범위  
           colors = pal)       # 색깔 목록
```

단어 색상 바꾸기

```
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 = 2,           # 최소 단어 빈도
           max.words = 200,        # 표현 단어 수
           random.order = F,       # 고빈도 단어 중앙 배치
           rot.per = .1,          # 회전 단어 비율
           scale = c(4, 0.3),     # 단어 크기 범위
           colors = pal)          # 색상 목록
```

실습

❖ 워드 클라우드의 각 옵션 수치를 변경해보자

국정원 트위터 텍스트 마이닝

국정원 트윗 텍스트 마이닝

- ❖ 국정원 대선 개입 사실이 밝혀져 논란이 됐던 2013년 6월, 독립 언론 뉴스타파가 인터넷을 통해 공개
- ❖ 국정원 계정으로 작성된 3,744개 트윗



국정원 대선개입에 분노해서 일어난 청소년들이 있었습니다 2020.02.2
2013년 7월, 청소년들은 국정원의 대선개입 사건에 분노해서 시국선언을
문화제를 열었습니다. #1. 엄재연(16·강원도 속초) "2008년 촛불집회 때 제
진보당 최서현 blog.naver.com/ujjuju595959/22182873385... | 블로그 내



국정원 대선개입 ... 2013년 10대뉴스 [2위], 2013.12.21.
2013년 10대 뉴스를 다루고 있는데요. 다음편은 '민영화'입니다. 1위 윤창
위 국정원 대선개입 3위 철도,의료 민영화 논란 [다음편] ...
돌고래 힐링스토리 blog.naver.com/kwcheal/140203167969 | 블로그 내



'국정원 대선개입' 시국촛불집회 - 2013년 6월 29일 2013.06.30.
적어도 '국정원 대선개입'이 명백하게 밝혀지기 전까지만 해도 말이다. 그
2012년12월19일 아침 공기를 가르며 투표하러 갈때만 해도 선거 막바지에.
행복은 뒤 돌아볼때... inmyframe.net/10171342796 | 블로그 내 검색

데이터 준비

```
# 데이터 로드
twitter <- read.csv("twitter.csv",
                    header = T,
                    stringsAsFactors = F,
                    fileEncoding = "UTF-8")

#내용 분석
str(twitter)
# 한글이 문자가 될 수 있으므로 변수명 수정
twitter <- rename(twitter,
                  no = 번호,
                  id = 계정이름,
                  date = 작성일,
                  tw = 내용)

#str_replace_all을 사용하기 위해 등록
library(stringr)
# 특수문자 제거
twitter$tw <- str_replace_all(twitter$tw, "WWW", " ")
```

단어 빈도표 만들기

```
#분석 데이터 확인인  
head(twitter$tw)  
# 트윗에서 명사추출  
nouns <- extractNoun(twitter$tw)  
# 추출한 명사 list를 문자열 벡터로 변환, 단어별 빈도표 생성  
wordcount <- table(unlist(nouns))  
# 데이터 프레임으로 변환  
df_word <- as.data.frame(wordcount, stringsAsFactors = F)
```


데이터 가공

```
# 변수명 수정
df_word <- rename(df_word,
                  word = Var1,
                  freq = Freq)
# 두 글자 이상 단어만 추출
df_word <- filter(df_word, nchar(word) >= 2)
# 상위 20개 추출
top20 <- df_word %>%
  arrange(desc(freq)) %>%
  head(20)
```

막대 그래프

```
library(ggplot2)
```

```
order <- arrange(top20, freq)$word          # 빈도 순서 변수 생성
```

```
ggplot(data = top20, aes(x = word, y = freq)) +  
  ylim(0, 2650) +  
  geom_col() +  
  #가로 막대로 그래프 사용  
  coord_flip() +  
  # 빈도 순서 변수 기준 막대 정렬  
  scale_x_discrete(limit = order) +          # 막대 그래프에 빈도 표시  
  geom_text(aes(label = freq), hjust = -0.3)
```

워드 클라우드

```
pal <- brewer.pal(8,"Dark2")    # 색상 목록 생성
set.seed(1234)                  # 난수 고정

wordcloud(words = df_word$word,  # 단어
  freq = df_word$freq,          # 빈도
  min.freq = 10,                 # 최소 단어 빈도
  max.words = 200,               # 표현 단어 수
  random.order = F,              # 고빈도 단어 중앙 배치
  rot.per = .1,                  # 회전 단어 비율
  scale = c(6, 0.2),             # 단어 크기 범위
  colors = pal)                  # 색상 목록
```