텍스트 마이닝

빅데이터 활용사례(1)

(1) 대형 유통 업체 타겟(Target)

(2) 미국의 월마트(WalMart)

빅데이터 활용사례(2)

(3) IBM 왓슨 – 환자 진단 및 치료 가이드 라인

(4) 아마존(Amazon)은 고객의 상품 데이터 분석 후 특정 상품을 구매한 사람이 추가로 구매할 것으로 예상되는 추천 시스템을 개발 매년 20%이상의 신장세

워드 클라우드(word cloud)

A. 텍스트에서 빈번히 사용된 키워드를 시각적으로 표시하는 텍스트 마이닝 방법.

B. 사용빈도가 높은 단어일수록 큰 글씨로 표시한다.

소스 코드(1)

```
library(KoNLP)
library(wordcloud)

text = "텍스트마이닝은 자연어(natural language)로 구성된 비정형 텍스트 데이터에서
패턴 또는 관계를 추출하여 가치와 의미 있는 정보를 찾아내는 마이닝 기법이다.
(나) 자연어 처리(natural language processing)기술에 기반한 방법이다.
텍스트마이닝은 말 그대로 텍스트 형태의 비정형 데이터에 마이닝 기법을 적용한 것이다. 즉 텍스트에
나타나는 단어를 분해, 정제하고, 특정 단어의 출현빈도 등을 파악하여 단어들 간의 관계를 조사하는
기법이다.
데이터마이닝은 대규모 DB에 저장된 정형화된 데이터로부터 정보를 찾아내는 기법이라면
텍스트마이닝은 비정형화된 텍스트 문서에서 정보를 찾아내는 기법이라고 할 수 있다."
useSejongDic()
nouns <- extractNoun(text)

nouns <- nouns[nchar(nouns) >=2 ]
nouns <- gsub("텍스트마이닝.*", "텍스트마이닝", nouns)
nouns <- gsub("데이터마이닝.*", "데이터마이닝", nouns)
nouns
```

gsub(pattern, replacement, x)

소스 코드(2)

```
wordFreq <- table(nouns)
pal <- brewer.pal(6,"Dark2")

windowsFonts(malgun=windowsFont("맑은 고딕"))

set.seed(1000)
wordcloud(words=names(wordFreq),
          freq=wordFreq,
          colors=pal,
          min.freq=1,
          random.order=F,
          family="malgun")
```

실행 결과


처리(natural
출현
데이터마이닝
language
특정
조사
텍스트마이닝
텍스트
의미
정제
비정형
빈도
마이닝
DB
방법
구축
관계
오정
다어
정보
기반
형태
기술
데이터
비정
패턴
적용
추출
processing
문서
대규모
저장
자연어(natural
파악

알아보기

(가) KoNLP 패키지

(나) useSejongDic() 함수

(다) extractNoun( )

(라) nchar( )

(마) gsub( )

(바) brewer.pal(n, name)

n : 워드 클라우드에서 색상의 수
name : 컬러 팔레트의 이름
Accent, Dark2, Paired, Pastel1, Pastel2, Set1, Set2, Set3
display.brewer.all( )

(사) windowFonts(malgun=windowsFont("맑은 고딕"))

wordcloud

```
wordcloud(words,
          freq,
          scale,
          min.freq,
          colors,
          random.order=TRUE,
          family, ...
```

word : 단어들의 이름(names)
freq : 단어들의 빈도(frequency)
scale : 가장 빈도 높은 단어의 크기와 가장 빈도 낮은 단어의 크기 지정
min.freq : 최소 빈도수 지정
colors : 단어의 색 지정
family : 폰트 지정
random.order =TRUE
=> 단어빈도와 상관없이 단어들이 나타난다.