14

비정형데이터분석

텍스트 데이터 분석 사례(1)

통계·데이터과학과 장영재 교수

학습목차

1. 분석대상텍스트데이터
2. 텍스트데이터의 전처리
3. 텍스트데이터의 탐색적자료분석

01

분석 대상 텍스트 데이터

1. 분석 대상 텍스트 데이터

1 로이터 코퍼스

- 로이터 코퍼스에는 RCV-1(Lewis et al., 2004), RCV-2 등 몇 가지 종류가 있음
  - R의 tm.corpus.Reuters21578 패키지(Theußl et al., 2012)에 포함되어 있는 Reuters-21578(Lewis et al., 1997)를 분석용 텍스트 데이터로 사용

→ 1987년 영문 기사 21,578건을 모아놓은 것으로서 기사 내용과 함께 기사에 대한 메타데이터, 즉 작성일자, 기사작성자, 주제의 범주 등에 대한 내용들도 수록

1. 분석 대상 텍스트 데이터

2 R을 이용한 텍스트 데이터 가공

- Reuters-21578 코퍼스를 불러오기 위해서는 "XML" 패키지와 "tm.corpus.Reuters21578" 패키지가 필요
  - data( ) 함수를 이용하여 코퍼스 Reuters21578를 불러올 수 있음

→ Reuters21578은 기사 내용과 기사에 대한 메타데이터들로 구성된 리스트들을 원소로 가지는 리스트 형태의 데이터

```
> install.packages("XML")
> install.packages("tm") # tm 패키지 사용이 어렵다는 경고 메시지 출력 시 설치 필요
> install.packages("tm.corpus.Reuters21578", repos = "http://datacube.wu.ac.at")
> library(tm.corpus.Reuters21578)
> data(Reuters21578)
```

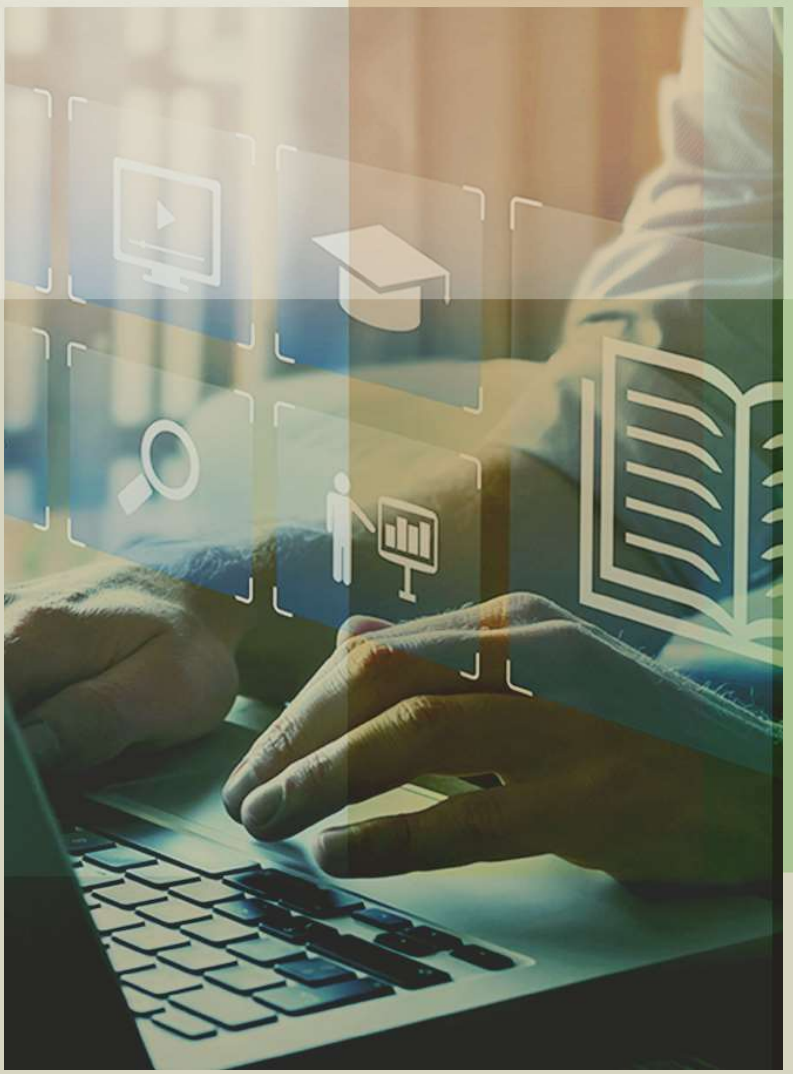
1. 분석 대상 텍스트 데이터

- 기사 내용과 메타데이터들의 리스트를 벡터로 변환하고 Reut_lists라는 이름으로 저장

```
> Reut_lists <- lapply(Reuters21578, FUN = unlist)
> names(Reut_lists[[1]])
> Reut_content <- lapply(Reut_lists, function(x) x[names(x) == "content"])
# 기사 내용 추출
> Reut_topics <- lapply(Reut_lists, function(x) x[names(x) == "meta.topics_cat"])
# 주제 추출
> sort(table(unlist(Reut_topics)), decreasing = T)
> Reut_content <- Reut_content[Reut_topics=="money-fx" | Reut_topics=="interest"]
# 외환과 금리 기사
> Reut_topics <- Reut_topics[Reut_topics=="money-fx" | Reut_topics=="interest"]
# 외환과 금리 기사
```

02

텍스트 데이터의 전처리



2. 텍스트 데이터의 전처리

- 로이터 코퍼스의 기사들에 대한 전처리 과정을 진행
  - 줄바꿈 기호 "₩n"가 다수 포함되어 있으므로 gsub( ) 함수로 줄바꿈 기호를 공백으로 변환하고 "'s" 삭제, 아포스트로피(')와 줄표(-)를 제외한 문장부호 삭제, 대소문자변환 작업을 수행하고 공백 문자 " "를 기준으로 기사 내용을 분할

```
> Reut_content <- gsub(Reut_content, pattern = "₩n", replacement = " ")
> Reut_content <- gsub(Reut_content, pattern = "'s", replacement = "")
> Reut_content <- gsub(Reut_content, pattern = "([^[:alnum:][:blank:]'-])", replacement = "")
> Reut_content <- tolower(Reut_content)
> Reut_content <- strsplit(Reut_content, " ")
```

2. 텍스트 데이터의 전처리

- 생성된 기사들의 리스트 Reut_content에는 내용이 누락된 기사들이 있으므로 이를 찾아 제거하고 나머지 기사에 불용어와 아포스트로피(')를 삭제하고 원형복원 작업을 수행

```
> which(Reut_content=="character0")
> Reut_topics <- Reut_topics[Reut_content!="character0"]
> Reut_content <- Reut_content[Reut_content!="character0"]
> library(stopwords)
> Reut_content <- lapply(Reut_content, function(x) x[! x %in% c(stopwords(), "")])
> Reut_content <- lapply(Reut_content, function(x) gsub(x, pattern = "'", replacement = ""))
> library(textstem)
> Reut_content <- lapply(Reut_content, lemmatize_strings)
```

2. 텍스트 데이터의 전처리

- 전체 단어 목록을 작성하기 위해 unlist( ) 함수로 모든 기사를 하나의 벡터로 묶고 unique( ) 함수로 이 벡터 내에 포함되어 있는 단어들을 중복되지 않게 선택하여 벡터를 생성

```
> Reut_lev <- sort(unique(unlist(Reut_content)))
> Reut_DTM <- lapply(Reut_content, FUN = function(x, lev){table(factor(x, lev, ordered
= T))}, lev = Reut_lev )
> Reut_DTM <- matrix(unlist(Reut_DTM), nrow = length(Reut_DTM), byrow = TRUE)
```

2. 텍스트 데이터의 전처리

- Reut_DTM 행렬은 470건의 기사 5,014개의 고유한 단어 포함

```
> dim(Reut_DTM)
[1]  470 5014
> sum(Reut_DTM>0)  # 0 아닌 셀
[1] 33248
> sum(Reut_DTM==0) # 0인 셀
[1] 2323332
> 1-sum(Reut_DTM>0)/sum(Reut_DTM>=0) # 희소도
[1] 0.9858914
> colnames(Reut_DTM) <- Reut_lev # Reut_DTM 각 열을 단어로
```

03

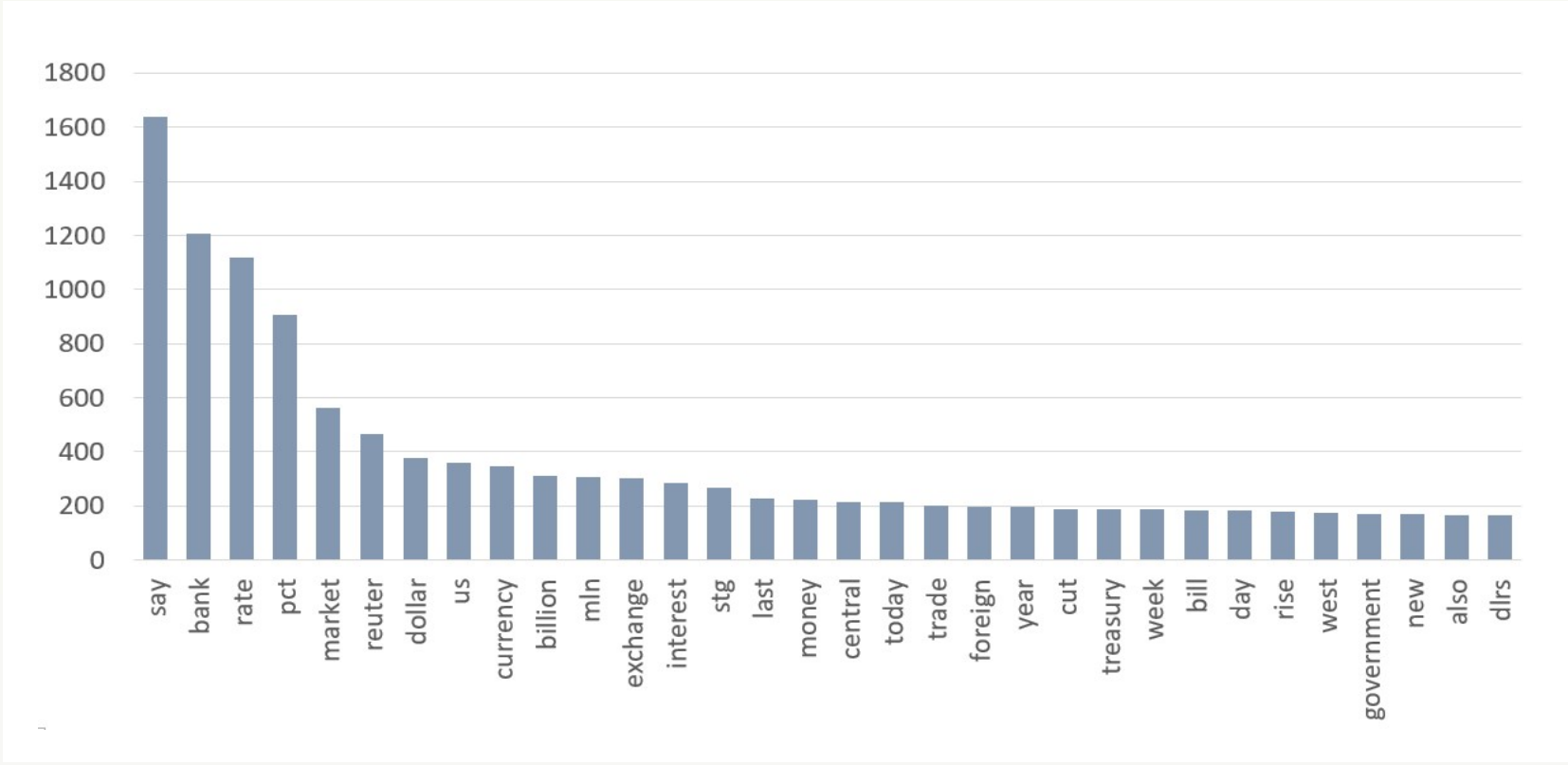
텍스트 데이터의 탐색적 자료분석

3. 텍스트 데이터의 탐색적 자료분석

1 전체 데이터에 대한 탐색적 자료분석

- 텍스트 데이터에서 전반적으로 빈번하게 사용된 단어들을 확인하기 위한 분석 절차 수행
  - table( ) 함수를 적용하여 도수분포표를 작성하고 출현빈도가 높은 단어들부터 확인하기 위해 sort( ) 함수를 이용하여 정렬

```
> Reut_table <- sort(table(unlist(Reut_content)), decreasing = T)
> Reut_table
> barplot(Reut_table[1:32])
```

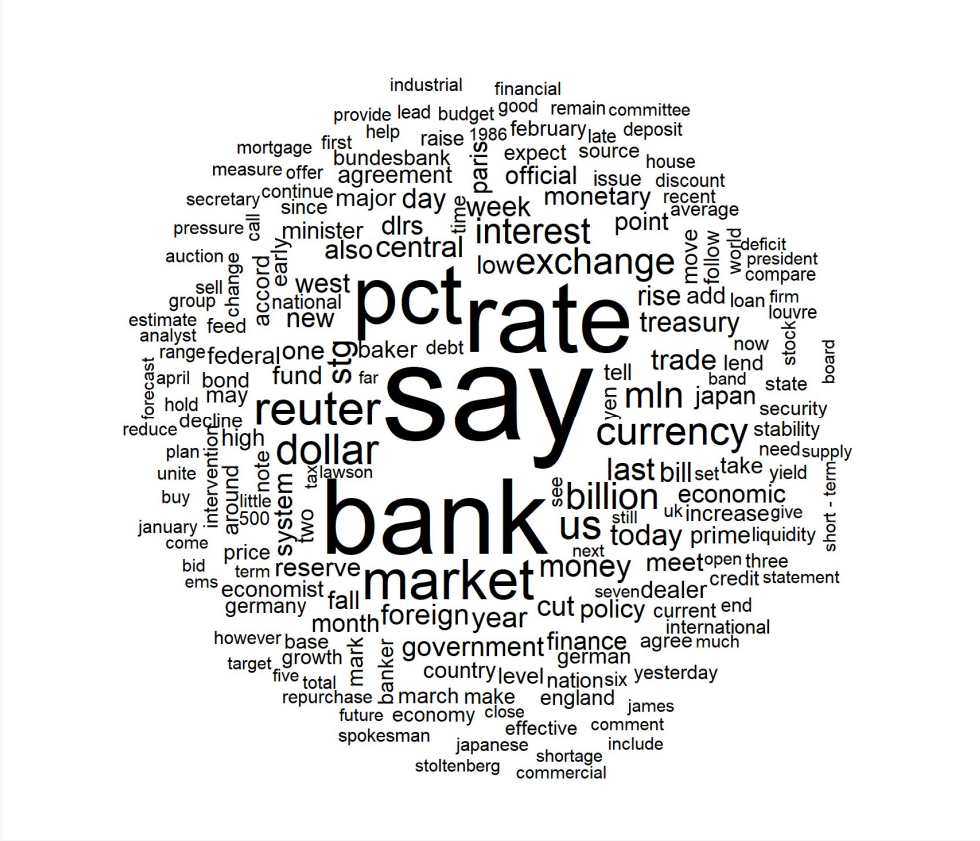


<그림> 외환과 금리기사 빈출 어휘

3. 텍스트 데이터의 탐색적 자료분석

- 도수분포표 Reut_table에 R wordcloud 패키지의 wordcloud( ) 함수를 적용하여 워드클라우드 작성

```
> library(wordcloud)
> wordcloud(words = names(Reut_table), freq = Reut_table,  max.words = 200,
random.order = F)
```


industrial
financial
provide
lead
budget
good
remain
committee
help
raise
1986
february
late
deposit
mortgage
first
paris
expect
source
bundesbank
house
measure
offer
agreement
official
issue
discount
secretary
continue
since
major
day
week
monetary
recent
average
pressure
call
minister
dlrs
time
interest
point
auction
early
also
central
low
exchange
move
follow
world
deficit
president
compare
sell
change
accord
west
pct
rate
rise
add
loan
firm
louvre
group
national
new
treasury
estimate
feed
analyst
range
federal
one
stg
baker
debt
now
stock
board
forecast
april
bond
fund
far
tell
trade
lend
band
state
hold
may
reuter
say
yen
mln
japan
security
reduce
decline
currency
stability
plan
high
dollar
need
supply
unite
intervention
around
note
tax
lawson
last
bill
set
take
yield
short - term
buy
system
two
bank
see
billion
economic
little
500
still
uk
increase
give
january
us
today
prime
liquidity
come
price
next
bid
term
reserve
market
money
meet
open
three
ems
economist
fall
seven
dealer
credit
statement
germany
foreign
year
cut
policy
current
end
international
however
base
month
government
finance
agree
much
target
growth
mark
banker
german
five
total
country
level
nation
six
yesterday
repurchase
march
make
england
james
future
economy
close
effective
comment
spokesman
japanese
include
shortage
stoltenberg
commercial

3. 텍스트 데이터의 탐색적 자료분석

2 문서 주제별 탐색적 자료분석

- 기사를 주제별로 나누어 탐색적 자료분석을 실시
  - 전체 코퍼스에서 주제가 "money-fx"인 기사들과 "interest"인 기사들을 Reut_fx과 Reut_int라는 리스트로 저장

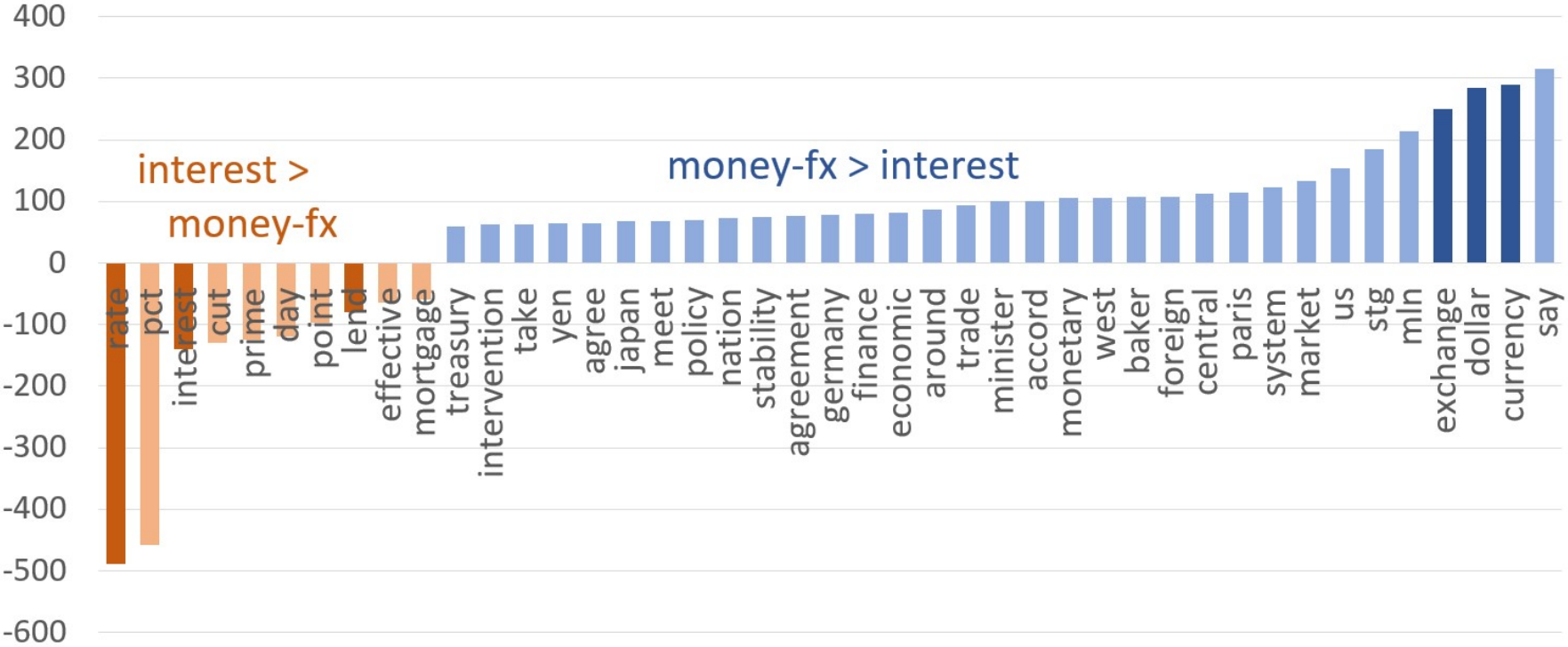
→ unlist( ) 함수로 두 리스트를 벡터로 전환한 후 table( ) 함수를 이용하여 도수분포표를 작성

```
> Reut_fx <- Reut_content[Reut_topics=="money-fx"]
> Reut_int <- Reut_content[Reut_topics=="interest"]
> Reut_fx_tab <- table(factor(unlist(Reut_fx), levels = Reut_lev, ordered = T))
> Reut_int_tab <- table(factor(unlist(Reut_int), levels = Reut_lev, ordered = T))
```

3. 텍스트 데이터의 탐색적 자료분석

→ 두 도수분포표의 차이를 구하여 Reut_fx_int이라는 이름으로 저장한 후 두 테이블의 출현빈도수 차이가 60이상인 단어들만 선택하여 막대그래프로 시각화

```
> Reut_fx_int <- sort(Reut_fx_tab - Reut_int_tab)
> Reut_fx_int <- Reut_fx_int[abs(Reut_fx_int)>=60]
> barplot(Reut_fx_int)
```

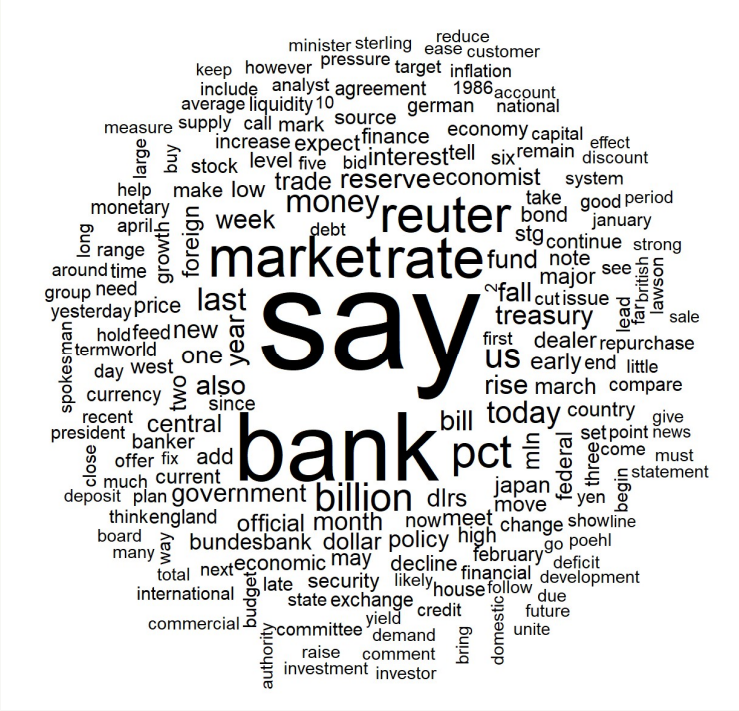


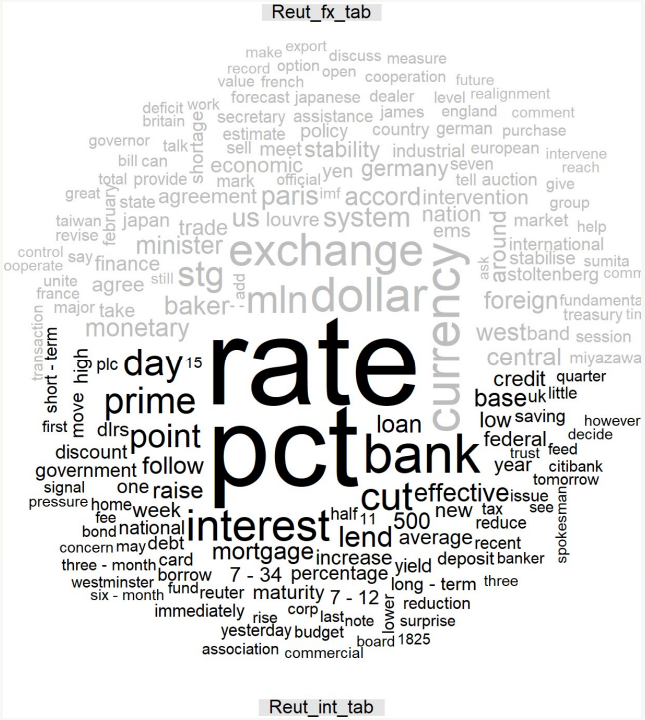
<그림> 주제별 출현빈도 차이

3. 텍스트 데이터의 탐색적 자료분석

→ 두 주제 기사의 공통되는 단어 클라우드(좌)와 공통되지 않는 단어 클라우드(우) 작성

```
> Reut_fx_int_mat <- cbind(Reut_fx_tab, Reut_int_tab)
> commonality.cloud(Reut_fx_int_mat, max.words = 200, random.order = FALSE)
> comparison.cloud(Reut_fx_int_mat, max.words = 200, random.order = FALSE,
colors = c("grey", "black"))
```

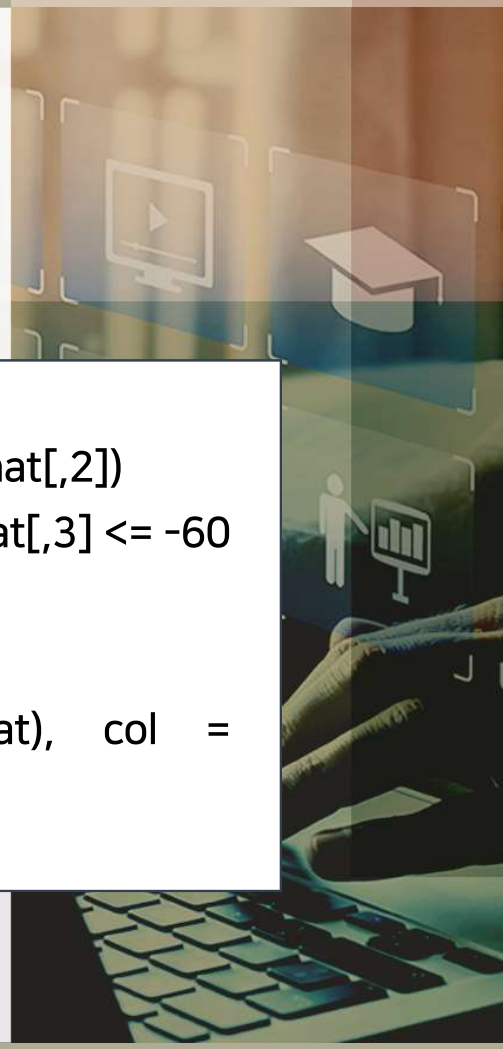

minister sterling reduce ease customer keep however pressure target inflation include analyst agreement 1986 account average liquidity 10 german national measure supply call mark source economy capital increase expect finance effect large buy stock level five bid interest tell six remain discount help make low trade reserve economist system monetary money reuter take bond good period april foreign week debt stg january range growth marketrate continue strong around time fund note group need major see yesterday price say fall cut issue british lawson last treasury lead far sale spokesman hold feed new year termworld one us early dealer repurchase day west end little currency two also rise march compare recent since today country give president central bill set point news banker add pct mln federal come must close offer fix three statement much current deposit plan government billion dlrs move japan yen begin think england official month now meet change showline board way bundesbank dollar policy high go poehl many total next economic may decline february deficit international late security likely financial development budget house follow due state exchange credit future commercial committee yield demand unite authority raise comment bring domestic investment investor

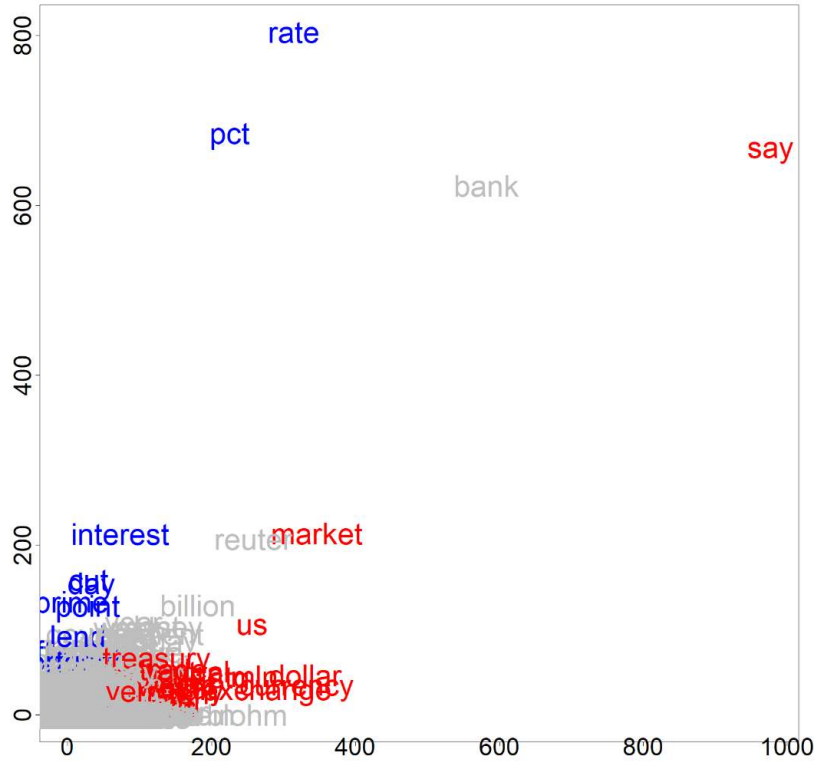

Reut_fx_tab
make export discuss measure record option open cooperation value french future deficit work forecast japanese dealer level realignment britain secretary assistance james england comment estimate policy country german purchase governor talk shortage sell meet stability industrial european bill can economic yen germany seven total provide mark official tell auction great state agreement paris imf accord intervention group february taiwan japan trade us louvre system nation market revise minister exchange ems around international control say finance stg mln dollar stabilise sumita operate unite agree still add astoltenberg france major take baker monetary foreign fundamenta transaction short - term move high plc day 15 rate currency central treasury tin first dlrs point prime pct bank loan credit quarter discount follow base uk little government one raise cut effective low saving however signal pressure home week interest lend federal decide fee national 500 new tax see bond concern may debt half 11 reduce trust feed three - month card mortgage increase average recent year citibank westminster borrow 7 - 34 percentage deposit banker tomorrow six - month fund reuter maturity 7 - 12 yield long - term three immediately rise corp last note lower reduction yesterday budget board 1825 surprise association commercial spokesman
Reut_int_tab

3. 텍스트 데이터의 탐색적 자료분석

→ 산점도로 자주 등장하는 단어들의 출현빈도를 비교. 두 주제에서의 출현빈도 차이를 구하여 빈도차에 따라 60이상(빨강 또는 파랑) 60미만 회색으로 표시 (x축: 외환, y축: 금리)

```
> Reut_fx_int_mat <- cbind(Reut_fx_int_mat, Reut_fx_int_mat[,1] - Reut_fx_int_mat[,2])
> Reut_fx_int_col <- ifelse(Reut_fx_int_mat[,3] >= 60, "red", ifelse(Reut_fx_int_mat[,3] <= -60
, "blue", "grey"))
> plot(Reut_fx_int_mat[,1], Reut_fx_int_mat[,2], type = "n", xlab = "", ylab = "")
> text(Reut_fx_int_mat[,1], Reut_fx_int_mat[,2], row.names(Reut_fx_int_mat), col =
Reut_fx_int_col)
```




rate
pct
say
bank
interest
reuter
market
billion
us
treasury
dollar
0
200
400
600
800
1000

실습하기

다음시간안내

15

텍스트 데이터 분석 사례(2)