text mining-4

20143311 김희수

A : Data Preparation

```
A1 <- read.csv("../commentscollect2.csv",
                  stringsAsFactors = F)

A1_An <- read.csv("../data/posts_An.csv")

A1_Y <- A1[A1$value == "<직원 인증>", ]
A1_N <- A1[A1$value == "<직원 미인증>", ]


A1$post <- gsub("<직.+증>","",A1$post)
```

B : Text Wording

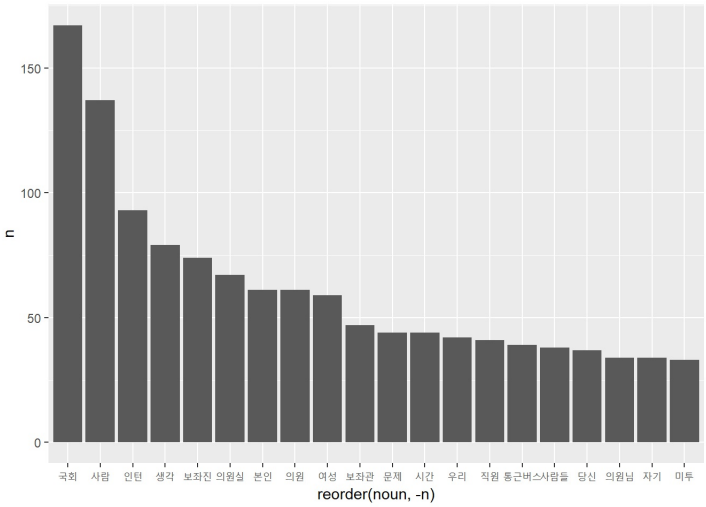
```
B1 <- A1$post %>%
  SimplePos09 %>%
  melt %>%
  as_tibble %>%
  select(3, 1)



#`명사에 관하여`
B2 <- B1 %>%
  mutate(noun=str_match(value, '([가-힣]+)/N')[,2]) %>%
  na.omit %>%
  filter(str_length(noun)>=2) %>%
  count(noun, sort=TRUE)

B2_2 <- head(B2,20)


ggplot(B2_2,aes(x=reorder(noun, -n),y=n)) +
  geom_bar(stat = "identity")
```

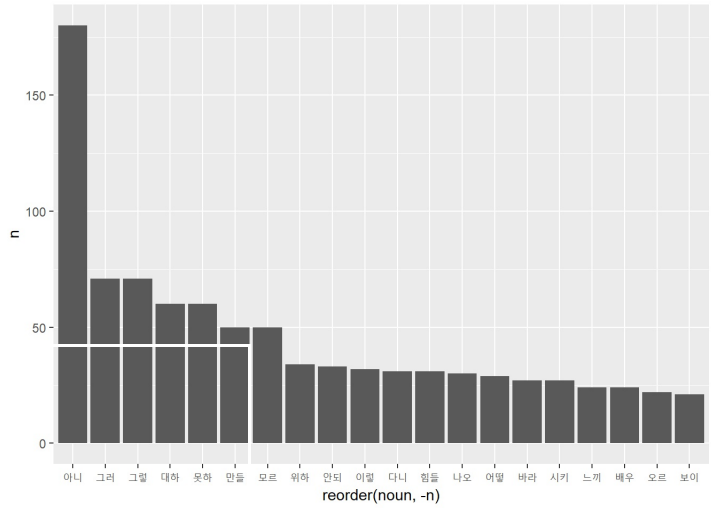



```
#'동사에 관하여'

B3 <- B1 %>%
  mutate(noun=str_match(value, '([가-힣]+)/P')[,2]) %>%
  na.omit %>%
  filter(str_length(noun)>=2) %>%
  count(noun, sort=TRUE)


B3_2 <- head(B3,20)


ggplot(B3_2,aes(x=reorder(noun, -n),
                y=n)) +
  geom_bar(stat = "identity")
```



국회:167

D : Word Cloud

```
wordcloud2(B2[1:30,],fontFamily = '나눔고딕',
             minRotation=0, maxRotation=0)
```

E : Co-Occurrence Network Anlysis

```
ko_words <- function(doc) {
  d <- as.character(doc)
  pos <- unlist(SimplePos22(d))

  extracted <- str_match(pos, '([가-힣]+)/[N][A-Z]')

  keyword <- extracted[, 2]
  keyword[!is.na(keyword)]
}

texts <- A1$post %>%
  str_replace_all(pattern="\r", replacement="") %>%
  str_replace_all(pattern="\n", replacement=" ") %>%
  str_replace_all(pattern="[[:punct:]]", replacement=" ") %>%
  str_replace_all(pattern="[ㄱ-ㅎㅏ-ㅣ]+", replacement="") %>%
  str_replace_all(pattern="/", replacement=" ") %>%
  str_trim(side="both")

texts <- texts[texts != ""]

pos <- Map(ko_words, texts)

corpus <- Corpus(VectorSource(pos))
tdm <- TermDocumentMatrix(corpus, control=list(
  removePunctuation=TRUE,
  removeNumbers=TRUE, wordLengths=c(4, 10), weighting=weightBin))
tdm.matrix <- as.matrix(tdm)

word.count <- rowSums(tdm.matrix)
word.order <- order(word.count, decreasing=TRUE)
freq.words <- tdm.matrix[word.order[1:100], ]


co.matrix <- freq.words %*% t(freq.words)
png("test.png",width=3.5,height=3.5,units="in",res=1200)
qgraph(co.matrix, labels=rownames(co.matrix),
       diag=FALSE, layout='spring', threshold=3,
       vsize=log(diag(co.matrix)) * 2)
```