1. "news_comment_BTS.csv"를 불러온 다음 행 번호를 나타낸 변수를 추가하고 분석에 적합하게 전처리하세요.

R 코드

```r
setwd('C:/Users/jspar/OneDrive/Documents/학교/전공/텍마')
raw_news_comment <- read_csv("news_comment_BTS.csv")
library(tidyverse)


#Q1.
news_comment <- raw_news_comment %>%
                select(reply) %>%
                mutate(reply = str_replace_all(reply, "[^가-힣]", " "),
                reply = str_squish(reply),
                id = row_number()) %>% print()
```

R 프로그램 결과

	reply	id
1	국보소년단	1
2	아줌마가 들어도 좋더라	2
3	팩트체크 현재 빌보드 위 방탄소년단 위 위 위 위 위 위 위...	3
4	방탄소년단이 한국사람이라 너무 자랑스러워요 우리오래오...	4
5	대단한 월드 클래스는 다르네 좋은 소식 응원해요	5
6	정국오빠 생일과 더불어 빌보드 위기사라니 축제구나	6
7	정말 축하하고 응원하지만 집에서 여러 계정으로 스트리밍 ...	7
8	기자는 자고 일어났지만 팬들은 못자고 발표 기다림	8
9	자랑스럽다 축하합니다	9
10	늘 응원하고 사랑합니다	10

Showing 1 to 10 of 1,200 entries, 2 total columns

2. 댓글에서 명사, 동사, 형용사를 추출하고 "/"으로 시작하는 모든 문자를 "다"로 바꾸시오.

R 코드

```r
#Q2.
library(KoNLP)
comment_pos <- news_comment %>%
               unnest_tokens(input = reply, output = word,
               token = SimplePos22, drop = F) %>%
               separate_rows(word, sep = "[+]") %>% print()


comment_new <- comment_pos %>%
               separate_rows(word, sep = "[+]") %>%
               filter(str_detect(word, "/n|/pv|/pa")) %>%
               mutate(word = ifelse(str_detect(word, "/pv|/pa"),
                                    str_replace(word, "/.*$", "다"),
                                    str_remove(word, "/.*$"))) %>%
               filter(str_length(word) >= 2) %>%
               arrange(id) %>% print()
```

R 프로그램 결과

	reply	id	word
1	국보소년단	1	국보소년
2	아줌마가 들어도 좋더라	2	아줌마
3	아줌마가 들어도 좋더라	2	들다
4	아줌마가 들어도 좋더라	2	좋다
5	팩트체크 현재 벌보드 위 방탄소년단 위 위 위 위 위 위 위...	3	팩트체크
6	팩트체크 현재 벌보드 위 방탄소년단 위 위 위 위 위 위 위...	3	벌보드
7	팩트체크 현재 벌보드 위 방탄소년단 위 위 위 위 위 위 위...	3	방탄소년단
8	방탄소년단이 한국사람이라 너무 자랑스러워요 우리오래오...	4	방탄소년단
9	방탄소년단이 한국사람이라 너무 자랑스러워요 우리오래오...	4	한국사람
10	방탄소년단이 한국사람이라 너무 자랑스러워요 우리오래오...	4	자랑

Showing 1 to 10 of 7,539 entries, 3 total columns

3. 다음 코드를 이용하여 유사어를 통일한 다음 한 댓글이 하나의 행이 되도록 단어를 결합하시오.

R 코드

```
#Q3.
comment_new <- comment_new %>%
               mutate(word = case_when(str_detect(., "축하") ~ "축하",
                                       str_detect(., "방탄") ~ "방탄",
                                       str_detect(., "대단") ~ "대단",
                                       str_detect(., "자랑") ~ "자랑",
                                       TRUE ~ word))

# 한 댓글이 하나의 행이 되도록 결합
line_comment <- comment_new %>%
                group_by(id) %>%
                summarise(sentence = paste(word, collapse = " ")) %>% print()
```

R 프로그램 결과

	id	sentence
1	1	국보소년
2	2	아줌마 들다 좋다
3	3	팩트체크 빌보드 방탄
4	4	방탄 한국사람 자랑 우리오래오래 함께하다
5	5	대단 월드 클래스 다르다 좋다 소식 응원해
6	6	정국오빠 생일 더불다 빌보드 위기사 축제구
7	7	축하 응원하지 계정 스트리밍 돌리다 사재기 팬덤 테러하 개...
8	8	기자 자다 일어나다 패다 못자 발표
9	9	자랑 축하
10	10	응원 사랑합

Showing 1 to 10 of 1,155 entries, 2 total columns

4. 댓글을 바이그램으로 토큰화 한 다음 바이그램 단어쌍을 분리하시오.

R 코드

```
#Q4.
# 바이그램 생성
bigram_comment <- line_comment %>%
                  unnest_tokens(input = sentence,
                                output = bigram,
                                token = "ngrams",
                                n = 2) %>% print()
# 바이그램 분리
bigram_seprated <- bigram_comment %>%
                   separate(bigram, c("word1", "word2"), sep = " ") %>% print()
```

R 프로그램 결과

- bigram_comment

	id	bigram
1	1	NA
2	2	아줌마 들다
3	2	들다 좋다
4	3	택트체크 벌보드
5	3	벌보드 방탄
6	4	방탄 한국사람
7	4	한국사람 자랑
8	4	자랑 우리오래오래
9	4	우리오래오래 함께하다
10	5	대단 월드

Showing 1 to 10 of 6,541 entries, 2 total columns

- bigram_seperated

	id	word1	word2
1	1	NA	NA
2	2	아줌마	들다
3	2	들다	좋다
4	3	택트체크	벌보드
5	3	벌보드	방탄
6	4	방탄	한국사람
7	4	한국사람	자랑
8	4	자랑	우리오래오래
9	4	우리오래오래	함께하다
10	5	대단	월드

Showing 1 to 10 of 6,541 entries, 3 total columns

5. 단어쌍 빈도를 구한 다음 네트워크 그래프 데이터를 생성하시오.

R 코드

```r
#Q5.
set.seed(1234)
pair_bigram <- bigram_seprated %>%
               count(word1, word2, sort = T) %>%
               na.omit() %>% print()

graph_bigram <- pair_bigram %>%
                filter(n >= 3) %>% # 적어도 빈도가 3이상만 선택
                as_tbl_graph(directed = F) %>%
                mutate(centrality = centrality_degree(),
                group = as.factor(group_infomap())) %>% print()
```

R 프로그램 결과

- pair_bigram

	word1	word2	n
1	축하	하다	43
2	방탄	축하	31
3	진짜	대단	23
4	자랑	방탄	21
5	축하	방탄	21
6	방탄	진짜	15
7	대단	자랑	14
8	방탄	자랑	14
9	빌보드	축하	14
10	진짜	자랑	14

Showing 1 to 10 of 5,504 entries, 3 total columns

- graph_bigram

```
# A tbl_graph: 88 nodes and 131 edges
#
# An undirected multigraph with 7 components
#
# Node Data: 88 x 3 (active)
  name   centrality group
  <chr>       <dbl> <fct>
1 축하             20 3
2 방탄             27 4
3 진짜             13 5
4 자랑             20 6
5 대단             11 9
6 빌보드           16 1
# ... with 82 more rows
#
# Edge Data: 131 x 3
   from    to     n
  <int> <int> <int>
1     1    61    43
2     1     2    31
3     3     5    23
# ... with 128 more rows
```

6. 바이그램을 이용하여 네트워크 그래프를 만드시오.

R 코드

```
#Q6.
library(showtext)
font_add_google(name = "Nanum Gothic", family = "nanumgothic")
showtext_auto()
set.seed(1234)
graph_bigram %>% ggraph(layout = "fr") + # 레이아웃
                 geom_edge_link(color = "gray50", # 엣지 색깔
                                alpha = 0.5) + # 엣지 명암
                 geom_node_point(aes(size = centrality, # 노드 크기
                                     color = group), # 노드 색깔
                                 show.legend = F) + # 범례 삭제
                 scale_size(range = c(4, 8)) + # 노드 크기 범위
                 geom_node_text(aes(label = name), # 텍스트 표시
                                repel = T, # 노드밖 표시
                                size = 5, # 텍스트 크기
                                family = "nanumgothic") + # 폰트
                 theme_graph()

install.packages('ggraph')
```

R 프로그램 결과