1. speech_park.txt를 불러와 분석에 맞게 전처리한 다음 띄어쓰기 기준으로 토큰화 하세요.

R 코드

```
```{r 과제_4, include=TRUE, echo=TRUE}
knitr::opts_chunk$set(echo = TRUE)

# 4-(1)
library(tidytext)
a <- readLines("speech_park.txt", encoding = "UTF-8")
speech <- tibble(value = a)
word_token <- speech %>% unnest_tokens(input = value, output= word, token = 'words')
word_token
```

R 프로그램 결과

Filter

	word
1	존경하는
2	국민
3	여러분
4	저는
5	오늘
6	국민
7	한
8	분
9	한
10	분의
11	꿈이
12	이루어지는
13	행복한

Showing 1 to 14 of 1,430 entries, 1 total columns
```

2. 가장 자주 사용된 단어 20개를 추출하세요.

R 코드

```
# 4-(2)
word_count <- word_token %>%
                count(word, sort = T) %>%
                filter(str_length(word) > 1)%>%
                head(20)
word_count
```

R 프로그램 결과

	word	n
1	국민	29
2	저는	14
3	있습니다	12
4	함께	12
5	꿈을	10
6	것입니다	8
7	새로운	8
8	있는	8
9	국민행복의	7
10	길을	7
11	것이	6
12	국민들의	6
13	만들겠습니다	6
14	박근혜	6
15	아니라	6
16	여러분의	6
17	우리	6
18	있도록	6
19	통해	6
20	대한	5

Showing 1 to 20 of 20 entries, 2 total columns

3. 가장 자주 사용된 단어 20개의 빈도를 나타낸 막대 그래프를 만드세요. (폰트: 나눔고딕)

R 코드

```
# 4-(3)
word_count %>%
  ggplot(aes(x=reorder(word, n), y = n)) +
         geom_col(colour='blue', fill='skyblue') +
         coord_flip() +
         geom_text(aes(label = n), hjust = -0.4) +
         labs(title = '박근혜 대통령 출마 연설문 단어 빈도',
                      x = NULL, y = NULL) +
         theme_set(theme_gray(base_family = 'nanumgothic'))
```

R 프로그램 결과