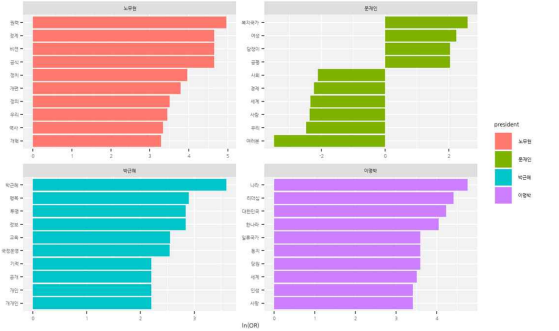
1. 4명의 대통령의 대선 출마 선언문의 명사를 추출하여 로그오즈비를 구하고 중요한 단어 10개씩을 뽑아서 막대그래프를 그리시오.

R 코드

```
#명사 토큰화 및 추출
speeches <- raw_speeches %>%
            mutate(value=str_replace_all(value, pattern="[^가-힣]", replacement=" "),
                   value=str_squish(value))

speeches_noun <- speeches %>%
                 unnest_tokens(input = value,
                               output = word,
                               token = extractNoun)

freq_noun <- speeches_noun %>%
             filter(str_length(word) > 1) %>% # 두 글자 이상 추출
             count(president, word) %>% # 연설문 및 단어별 빈도
             print()

library(tidylo)
freq_noun_lo = freq_noun %>%
               filter(word != "국민") %>%
               bind_log_odds(set = president, feature = word, n = n) %>%
               arrange(-log_odds_weighted) %>% # moon에서 비중이 큰 단어
               print()

top10_lo2 = freq_noun_lo %>%
            group_by(president) %>%
            slice_max(abs(log_odds_weighted), n=10, with_ties=FALSE) %>%
            arrange(-log_odds_weighted) %>% print(n=Inf)

top10_lo2 %>% ggplot(aes(x = reorder_within(x=word, by=log_odds_weighted, within=president),
                         y = log_odds_weighted, fill = president)) +
              geom_col() +
              coord_flip() +
              ylab("ln(OR)") +
              facet_wrap(~ president, scales = "free", ncol = 2) +
              scale_x_reordered() +
              labs(x = NULL) +
              theme(text = element_text(family = "nanumgothic"))
```

R 프로그램 결과



2. Speeches_presidents.csv를 불러와 이명박 전 대통령과 노무현 전 대통령의 연설문만을 추출하여 다음에 답하시오.

(1). 연설문에서 명사를 추출한 다음 연설문 별 단어 빈도를 구하세요.

R 코드

```
# 2-(1)
speeches <- raw_speeches %>%
            mutate(value=str_replace_all(value, pattern="[^가-힣]", replacement=" "),
                   value=str_squish(value))

speeches_noun <- speeches %>%
                 unnest_tokens(input = value,
                               output = word,
                               token = extractNoun)

freq_noun <- speeches_noun %>%
             count(president, word) %>%
             filter(str_length(word) > 1) %>%
             print()

freq_noun %>% filter(president == "이명박")
freq_noun %>% filter(president == "노무현")
```

R 프로그램 결과

president <chr>	word <chr>	n <int>
노무현	가슴	2
노무현	가톨	2
노무현	갈등	1
노무현	감옥	1
노무현	강자	1
노무현	개편	4
노무현	개혁	4
노무현	건국	1
노무현	경선	1
노무현	경쟁	1

1-10 of 216 rows

president	word <chr>	n <int>
이명박	가능	1
이명박	가족	1
이명박	가지	3
이명박	가치	2
이명박	각계각층	1
이명박	각국	1
이명박	감사	1
이명박	강국	1
이명박	개발	1
이명박	걱정	1

1-10 of 202 rows

2-(2). 로그RR를 이용해 두 연설문에서 상대적으로 중요한 단어를 10개씩 추출하세요.

R 코드

```
# 2-(2)
library(tidylo)
freq_noun_lo = freq_noun %>%
               bind_log_odds(set = president, feature = word, n = 10) %>%
               arrange(log_odds_weighted)

freq_noun_wide <- freq_noun %>%
                  pivot_wider(names_from = president,
                              values_from = n,
                              values_fill = 0) %>%
                  mutate(p_lee = ((이명박+1)/(sum(이명박+1))),
                         p_no = ((노무현+1)/(sum(노무현+1))),
                         RR = p_lee/p_no,  # 상대위험 (RR) 변수 추가
                         log_RR = log(RR)) %>%
                  print()

top10_logRR = freq_noun_wide %>%
              group_by(president = ifelse(log_RR > 0, "이명박", "노무현")) %>%
              slice_max(abs(log_RR), n=10, with_ties=FALSE) %>%
              arrange(-log_RR) %>%
              select(word, 이명박, 노무현, log_RR, president) %>%
              print(n=Inf)
```

R 프로그램 결과

	word	이명박	노무현	log_RR	president
1	나라	15	0	2.764039	이명박
2	대한민국	12	0	2.556399	이명박
3	세계	13	1	1.937360	이명박
4	러대십	6	0	1.937360	이명박
5	여러분	11	1	1.783210	이명박
6	국가	5	0	1.783210	이명박
7	발전	5	0	1.783210	이명박
8	사람	5	0	1.783210	이명박
9	인생	5	0	1.783210	이명박
10	따뜻	4	0	1.600888	이명박
11	개편	0	4	-1.617988	노무현
12	개혁	0	4	-1.617988	노무현
13	당당	0	4	-1.617988	노무현
14	정의	0	4	-1.617988	노무현
15	지역	0	4	-1.617988	노무현
16	공식	0	6	-1.954460	노무현
17	비전	0	6	-1.954460	노무현
18	정계	0	6	-1.954460	노무현
19	정치	0	8	-2.205774	노무현
20	권력	0	9	-2.311135	노무현

Showing 1 to 20 of 20 entries, 5 total columns

2-(3). 두 연설문에서 상대적으로 중요한 단어를 나타낸 막대 그래프를 만드세요.

R 코드

```r
# 2-(3)
top10_logRR %>% ggplot(aes(x = reorder(word, log_RR),
                           y = log_RR, fill = president)) +
                geom_col() +
                coord_flip() +
                ylab("ln(RR)") +
                labs(x = NULL) +
                theme(text = element_text(family = "nanumgothic"))
```

R 프로그램 결과