Q1. speeches_roh.csv를 불러온 다음 연설문이 들어있는 content를 문장 기준으로 토큰화하시오.

- R 프로그램

```r
```{r 텍마 과제 10, include=TRUE, echo=TRUE}

library(tidyverse)
setwd('C:/Users/jspar/OneDrive/Documents/학교/전공/텍마')
raw_speech <- read_csv("speeches_roh.csv")


#Q1.
# 문장 토큰화
library(tidytext)
library(KoNLP)
sentences_speech <- raw_speech %>%
                    unnest_tokens(input = content,
                                  output = word,
                                  token = 'sentences',
                                  drop = F) %>%
                    filter(str_length(word) > 1) %>%
                    rename(sentences = word)
```
```

# Q1. speeches_roh.csv를 불러온 다음 연설문이 들어있는 content를 문장 기준으로 토큰화하시오.

- R 결과

| | date | president | place | event | source | paragraph | content | id | sentences |
|---|---|---|---|---|---|---|---|---|---|
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 주한외교사절 여러분 그리고 멀리서 오신 외빈 여러분 전두... |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 저의 대통령 취임을 축하해 주기 위해 이 자리에 참석하신 여... |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 이제 막 대통령의 책무를 부여받은 저는 기쁨보다는 무거운 ... |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 참으로 어깨가 무겁습니다 |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 그러나 이 자리까지 저를 이끌어주신 국민 여러분을 믿고 이 ... |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 신명을 다 바쳐 국민 여러분의 기대에 부응해 나가겠습니다 |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 존경하는 내외귀빈 여러분 |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 지금 우리는 선택의 기로에 서 있습니다 |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 도약과 후퇴의 중대한 갈림길입니다 |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 우리 국민은 숱한 도전을 슬기롭게 극복하면서 여기까지 왔... |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 끊임없는 외침을 극복하며 민족의 자존을 지켜왔습니다 |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 식민통치와 전쟁의 폐허를 딛고일어서 우리 경제를 세계 열... |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 분단이라는 악조건을 이겨내면서 성숙한 민주주의를 이루어... |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 참으로 자랑스런 우리 국민이 아닐 수 없습니다 |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 오늘의 우리를 있게 한 선열과 우리 국민께 무한한 존경과 감... |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 그러나 아직 우리가 가야할 길에는 불안과 희망이 교차하고 ... |
| 축연 연설 | 2003.02.25 | 노무현 | 국내 | 취임사 | 노무현대통령연설문집 제1권 2월 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 세계경제의 각축을 지역해지고 있습니다 |

Showing 1 to 17 of 39,714 entries, 10 total columns

# Q2. 문장을 분석에 적합하게 전처리한 다음 명사를 추출하시오.

- R 프로그램

```r
#Q2.
# 전처리
sentences_speech <- sentences_speech %>%
                    mutate(sentences = str_replace_all(sentences, "[^가-힣]", " "),
                           sentences = str_squish(sentences))
# 명사추출
noun_speech <- sentences_speech %>%
               unnest_tokens(input = sentences,
                             output = word,
                             token = extractNoun,
                             drop = F) %>%
               filter(str_length(word) > 1)
```

# Q2. 문장을 분석에 적합하게 전처리한 다음 명사를 추출하시오.

- R 결과

| ource | paragraph | content | id | sentences | word |
|---|---|---|---|---|---|
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 주한외교사절 여러분 그리고 멀리서 오신 외빈 여러분 전두... | 주한외교사절 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 주한외교사절 여러분 그리고 멀리서 오신 외빈 여러분 전두... | 여러분 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 주한외교사절 여러분 그리고 멀리서 오신 외빈 여러분 전두... | 외빈 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 주한외교사절 여러분 그리고 멀리서 오신 외빈 여러분 전두... | 여러분 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 주한외교사절 여러분 그리고 멀리서 오신 외빈 여러분 전두... | 전두환 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 주한외교사절 여러분 그리고 멀리서 오신 외빈 여러분 전두... | 대통령 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 주한외교사절 여러분 그리고 멀리서 오신 외빈 여러분 전두... | 부요 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 주한외교사절 여러분 그리고 멀리서 오신 외빈 여러분 전두... | 비롯 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 주한외교사절 여러분 그리고 멀리서 오신 외빈 여러분 전두... | 내빈 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 저의 대통령 취임을 축하해 주기 위해 이 자리에 참석하신 여... | 대통령 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 저의 대통령 취임을 축하해 주기 위해 이 자리에 참석하신 여... | 취임 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 저의 대통령 취임을 축하해 주기 위해 이 자리에 참석하신 여... | 축하 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 저의 대통령 취임을 축하해 주기 위해 이 자리에 참석하신 여... | 자리 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 저의 대통령 취임을 축하해 주기 위해 이 자리에 참석하신 여... | 참석 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 저의 대통령 취임을 축하해 주기 위해 이 자리에 참석하신 여... | 하신 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 저의 대통령 취임을 축하해 주기 위해 이 자리에 참석하신 여... | 여러분 |
| c.무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 저의 대통령 취임을 축하해 주기 위해 이 자리에 참석하신 여... | 감사 |

Showing 1 to 17 of 275,776 entries, 11 total columns

# Q3. 연설문 내 중복 단어를 제거하고 빈도가 100회 이하인 단어를 추출하시오.

- R 프로그램

```r
#Q3.
# 중복단어 제거
count_word <- noun_speech %>%
              add_count(word, sort=TRUE) %>%
              filter(!word %in% c("우리", "대통령", "있습니다", "있습니", "여러분", "국민"))
# 단어 빈도 100회 이하 추출
word_count <- count_word %>%
              filter(n<=100)
```

# Q3. 연설문 내 중복 단어를 제거하고 빈도가 100회 이하인 단어를 추출하시오.

- R 결과

| | paragraph | content | id | sentences | word | n |
|---|---|---|---|---|---|---|
| 1 | 연설문집 제1권 2일 | 1 존경하는 국민 여러분. 오늘 저는 대한민국의 제16대 대통령... | 2 | 동북아 시대는 경제에서 출발합니다 | 출발 | 100 |
| 2 | 연설문집 제1권 3일 | 1 친애하는 해군사관학교 졸업생 여러분, 자리를 함께 하신 학... | 9 | 오늘이 있기까지 정성과 사랑으로 뒷받침해주신 학부모님께... | 뒷받침 | 100 |
| 3 | 연설문집 제1권 3일 | 1 친애하는 해군사관학교 졸업생 여러분, 자리를 함께 하신 학... | 9 | 육 해 공 군이 균형있게 발전해 나가고 과학군 정보군으로 한... | 자원 | 100 |
| 4 | 연설문집 제1권 3일 | 1 친애하는 공군사관학교 졸업생 여러분, 사관생도와 공군장... | 11 | 이제 여러분은 영광스런 대한민국 공군장교로 새롭게 출발... | 출발 | 100 |
| 5 | 연설문집 제1권 3일 | 1 존경하는 상공인 여러분, 이 자리에 참석하신 내외 귀빈 여러... | 12 | 기술혁신은 우수한 인재를 키우는 데서 출발합니다 | 출발 | 100 |
| 6 | 연설문집 제1권 3일 | 1 존경하는 상공인 여러분, 이 자리에 참석하신 내외 귀빈 여러... | 12 | 경제를 효율적으로 뒷받침하는 서비스 행정이 이루어져야 ... | 뒷받침 | 100 |
| 7 | 연설문집 제1권 3일 | 1 경찰대학 19기 졸업생 여러분, 학부모와 내외 귀빈 여러분, ... | 13 | 여러분의 앞길에 영광과 보람이 함께 하기를 기원하면서 다... | 출발 | 100 |
| 8 | 연설문집 제1권 4일 | 1 존경하는 국회의장, 그리고 국회의원 여러분, 이 자리를 지켜... | 16 | 아울러 뜻 있는 젊은이들이 친구나 친지들로부터 도움을 받... | 출발 | 100 |
| 9 | 연설문집 제1권 4일 | 1 존경하는 과학기술인 여러분, 오늘 제36회 '과학의 날'을 진... | 25 | 과학기술 입국은 우수한 인재를 키우는 데서 출발합니다 | 출발 | 100 |
| 10 | 연설문집 제1권 5일 | 1 존경하는 국민 여러분, 저는 오늘부터 17일까지 미국을 방문... | 31 | 저와 동행하는 우리 경제인들도 민간 차원에서 구체적인 협... | 차원 | 100 |
| 11 | 연설문집 제1권 5일 | 1 5년 전 김대중 대통령이 외환위기를 맞아 미국에 다녀갔습니... | 32 | 이번 방미에서도 성공을 위한 큰 뒷받침이 될 것이라고 생각... | 뒷받침 | 100 |
| 12 | 연설문집 제1권 5일 | 1 5년 전 김대중 대통령이 외환위기를 맞아 미국에 다녀갔습니... | 32 | 국가의 발전도 국민의 행복도 평화로부터 출발합니다 | 출발 | 100 |
| 13 | 연설문집 제1권 5일 | 1 5년 전 김대중 대통령이 외환위기를 맞아 미국에 다녀갔습니... | 32 | 많은 어려움을 극복하고 도전정신과 열정으로 오늘의 한국... | 뒷받침 | 100 |
| 14 | 연설문집 제1권 5일 | 1 존경하는 토머스 도노휴(Thomas Donohue) 美 상공회의소 ... | 34 | 노동운동이 부당하게 탄압 받을 때는 인권 수호 차원에서 노... | 차원 | 100 |
| 15 | 연설문집 제1권 5일 | 1 존경하는 국민 여러분, 저는 첫 미국방문과 한미 정상회담을 ... | 37 | 저의 활동을 직접 돕기도 하고 활발한 투자유치와 무역상담 ... | 뒷받침 | 100 |
| 16 | 연설문집 제1권 5일 | 1 오늘은 미합중국 전몰장병을 추모하는 뜻깊은 날입니다. ... | 40 | 나와 조지 부시 대통령은 지난주 정상회담에서 한미동맹을 ... | 차원 | 100 |
| 17 | 연설문집 제1권 5일 | 1 존경하는 배리 오키프(Barry O'Keefe) 뱁슨칼재단의 이장 | 41 | 그 출발점은 지난 시대의 잘못된 관행을 정상화시키는 것이 | 출발 | 100 |

Showing 1 to 17 of 110,156 entries, 12 total columns

# Q4. 추출한 단어에서 다음의 불용어를 제거하시오.

- R 프로그램

```r
#Q4.
# 불용어 리스트
stopword <- c( "들이", "하다", "하게", "하면", "해서", "이번", "하네",
               "해요", "이것", "니들", "하기", "하지", "한거", "해주",
               "그것", "어디", "여기", "까지", "이거", "하신", "만큼")
# 불용어 제거
word_count <- word_count %>%
              filter(!word %in% stopword)
```

# Q4. 추출한 단어에서 다음의 불용어를 제거하시오.

- R 결과

| | paragraph | content | id | sentences | word | n |
|---|---|---|---|---|---|---|
| 1 | 연설문집 제1권 2일 | 1 존경하는 국민 여러분, 오늘 저는 대한민국의 제16대 대통령... | 2 | 동북아 시대는 경제에서 출발합니다 | 출발 | 100 |
| 2 | 연설문집 제1권 3일 | 1 친애하는 해군사관학교 졸업생 여러분, 자리를 함께 하신 학... | 9 | 오늘이 있기까지 정성과 사랑으로 뒷받침해주신 학부모님께... | 뒷받침 | 100 |
| 3 | 연설문집 제1권 3일 | 1 친애하는 해군사관학교 졸업생 여러분, 자리를 함께 하신 학... | 9 | 육 해 공 군이 균형있게 발전해 나가고 과학군 정보군으로 한... | 차원 | 100 |
| 4 | 연설문집 제1권 3일 | 1 친애하는 공군사관학교 졸업생 여러분, 사관생도와 공군장... | 11 | 이제 여러분은 영광스런 대한민국 공군장교로 새롭게 출발... | 출발 | 100 |
| 5 | 연설문집 제1권 3일 | 1 존경하는 상공인 여러분, 이 자리에 참석하신 내외 귀빈 여러... | 12 | 기술혁신은 우수한 인재를 키우는 데서 출발합니다 | 출발 | 100 |
| 6 | 연설문집 제1권 3일 | 1 존경하는 상공인 여러분, 이 자리에 참석하신 내외 귀빈 여러... | 12 | 경제를 효율적으로 뒷받침하는 서비스 행정이 이루어져야 ... | 뒷받침 | 100 |
| 7 | 연설문집 제1권 3일 | 1 경찰대학 19기 졸업생 여러분, 학부모와 내외 귀빈 여러분, ... | 13 | 여러분의 앞길에 영광과 보람이 함께 하기를 기원하면서 다... | 출발 | 100 |
| 8 | 연설문집 제1권 4일 | 1 존경하는 국회의장, 그리고 국회의원 여러분, 이 자리를 지켜... | 16 | 아울러 뜻 있는 젊은이들이 친구나 친지들로부터 도움을 받... | 출발 | 100 |
| 9 | 연설문집 제1권 4일 | 1 존경하는 과학기술인 여러분, 오늘 제36회 '과학의 날'을 진... | 25 | 과학기술 입국은 우수한 인재를 키우는 데서 출발합니다 | 출발 | 100 |
| 10 | 연설문집 제1권 5일 | 1 존경하는 국민 여러분, 저는 오늘부터 17일까지 미국을 방문... | 31 | 저와 동행하는 우리 경제인들도 민간 차원에서 구체적인 협... | 차원 | 100 |
| 11 | 연설문집 제1권 5일 | 1 5년 전 김대중 대통령이 외환위기를 맞아 미국에 다녀갔습니... | 32 | 이번 방미에서도 성공을 위한 큰 뒷받침이 될 것이라고 생각... | 뒷받침 | 100 |
| 12 | 연설문집 제1권 5일 | 1 5년 전 김대중 대통령이 외환위기를 맞아 미국에 다녀갔습니... | 32 | 국가의 발전도 국민의 행복도 평화로부터 출발합니다 | 출발 | 100 |
| 13 | 연설문집 제1권 5일 | 1 5년 전 김대중 대통령이 외환위기를 맞아 미국에 다녀갔습니... | 32 | 많은 어려움을 극복하고 도전정신과 열정으로 오늘의 한국... | 뒷받침 | 100 |
| 14 | 연설문집 제1권 5일 | 1 존경하는 토머스 도노휴(Thomas Donohue) 美 상공회의소 ... | 34 | 노동운동이 부당하게 탄압 받을 때는 언권 수호 차원에서 노... | 차원 | 100 |
| 15 | 연설문집 제1권 5일 | 1 존경하는 국민 여러분, 저는 첫 미국방문과 한미 정상회담을 ... | 37 | 저의 활동을 직접 돕기도 하고 활발한 투자유치와 무역상담 ... | 뒷받침 | 100 |
| 16 | 연설문집 제1권 5일 | 1 오늘은 미합중국 전몰장병들을 추모하는 뜻깊은 날입니다. ... | 40 | 나와 조지 부시 대통령은 지난주 정상회담에서 한미동맹을 ... | 차원 | 100 |
| 17 | 연설문집 제1권 5일 | 1 존경하는 배리 오키프(Barry O'Keefe) 밥풀굴제단의 이장 ... | 41 | 그 출발점은 지난 시대의 잘못된 과행을 정상화시키는 것이 | 출발 | 100 |

Showing 1 to 17 of 110,037 entries, 12 total columns

# Q5. 연설문 별 단어 빈도를 구한 다음 DTM을 만드시오.

- R 프로그램

```r
#Q5.
# 연설문 별 단어 빈도
count_word_speech <- word_count %>%
                     count(id, word, sort = T)
# DTM 생성
dtm_comment <- count_word_speech %>%
               cast_dtm(document = id, term = word, value = n) %>% print()
# DTM의 내용 확인하기
as.matrix(dtm_comment[1:15, 1:15])
```

# Q5. 연설문 별 단어 빈도를 구한 다음 DTM을 만드시오.

- R 결과

| Docs | Terms 대학교 | 고등학교 | 여성 | 뉴질랜드 | 사상 | 법안 | 김용옥 | 기자 | 베트남 | 인도네시아 | 매체 | 자율 | 북쪽 | 본고사 | 이집트 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 671 | 47 | 33 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 24 | 0 | 23 | 0 | |
| 65 | 0 | 0 | 31 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 626 | 0 | 0 | 0 | 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 695 | 0 | 0 | 1 | 0 | 28 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | |
| 699 | 0 | 0 | 0 | 0 | 28 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 709 | 0 | 0 | 0 | 0 | 0 | 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 180 | 0 | 1 | 0 | 0 | 0 | 1 | 25 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | |
| 698 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 264 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 621 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 24 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 650 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 2 | 0 | 0 | 24 | 0 | 0 | 0 | 0 | 0 | |
| 758 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 24 | 0 | 0 | |
| 507 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 22 | |
| 541 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | |
| 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | |

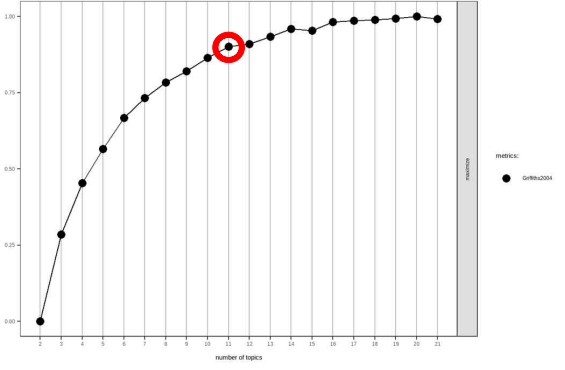
# Q6. 토픽 수를 2~20개로 바꿔가며 LDA 모델을 만든 다음 최적 토픽 수를 구하시오.

- R 프로그램

```r
#Q6.
# 토픽 수를 2~20개로 바꿔가며 LDA 모델을 만든 다음 최적 토픽 수를 구하시오
library(ldatuning)
models <- FindTopicsNumber(dtm = dtm_comment,
                           topics = 2:21, # 22개의 LDA 모델을 만듬
                           return_models = T,
                           control = list(seed = 1234))
models %>% select(topics, Griffiths2004)
FindTopicsNumber_plot(models)
```

# Q6. 토픽 수를 2~20개로 바꿔가며 LDA 모델을 만든 다음 최적 토픽 수를 구하시오.

- R 결과




• 11부터 비교적 완만한 곡선의 양상을 보임.
-> 최적 토픽 수: 11

## Q7. 토픽 수가 9개인 LDA 모델을 추출하세요.

- R 프로그램

```r
#Q7.
# 토픽 수 9개인 LDA 모델 생성
library(topicmodels)
lda_model <- LDA(dtm_comment,
                 k = 9,
                 method = "Gibbs",
                 control = list(seed = 1234))
# 모델 내용 확인
glimpse(lda_model)
```

# Q7. 토픽 수가 9개인 LDA 모델을 추출하세요.

- R 결과

```
Formal class 'LDA_Gibbs' [package "topicmodels"] with 16 slots
  ..@ seedwords      : NULL
  ..@ z              : int [1:110037] 1 1 1 1 1 1 1 1 1 1 ...
  ..@ alpha          : num 5.56
  ..@ call           : language LDA(x = dtm_comment, k = 9, method = "Gibbs", control = list(seed =
1234))
  ..@ Dim            : int [1:2] 780 16598
  ..@ control        :Formal class 'LDA_Gibbscontrol' [package "topicmodels"] with 14 slots
  ..@ k              : int 9
  ..@ terms          : chr [1:16598] "대학교" "고등학교" "여성" "뉴질랜드" ...
  ..@ documents      : chr [1:780] "671" "65" "626" "695" ...
  ..@ beta           : num [1:9, 1:16598] -4.95 -12 -11.63 -12.33 -12.04 ...
  ..@ gamma          : num [1:780, 1:9] 0.743 0.0328 0.0952 0.0354 0.018 ...
  ..@ wordassignments:List of 5
  .. ..$ i    : int [1:81871] 1 1 1 1 1 1 1 1 1 1 ...
  .. ..$ j    : int [1:81871] 1 2 9 12 14 17 21 31 33 37 ...
  .. ..$ v    : num [1:81871] 1 1 3 1 1 1 1 4 1 1 ...
  .. ..$ nrow: int 780
  .. ..$ ncol: int 16598
  .. ..- attr(*, "class")= chr "simple_triplet_matrix"
  ..@ loglikelihood  : num -864226
  ..@ iter           : int 2000
  ..@ logLiks        : num(0)
  ..@ n              : int 110037
```

Q8. LDA 모델의 beta를 이용해 각 토픽에 등장할 확률이 높은 상위 10개 단어를 추출한 다음 토픽 별 주요 단어를 나타낸 막대 그래프를 만드시오.

- R 프로그램

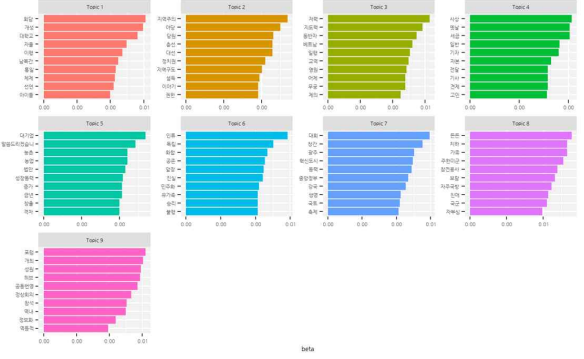
```r
#Q8.
# 단어들이 토픽 별로 들어갈 확률 베타 포함한 df
library(reshape2)
term_topic <- tidy(lda_model, matrix = "beta") %>%
              mutate(topic_name = paste("Topic", topic)) %>% # 토픽 이름 변수 설정
              print()
# beta 내용 확인
# 토픽 별 단어 수
term_topic %>% count(topic)

# 토픽별 beta 상위 10개 단어 추출
top_term_topic <- term_topic %>%
                  group_by(topic) %>%
                  slice_max(beta, n = 10, with_ties =F) %>% print() # 동점 제외
# 막대그래프 작성
library(showtext)
font_add_google(name = "Nanum Gothic", family = "nanumgothic")
showtext_auto() #한글폰트설정

library(scales)
top_term_topic %>% ggplot(aes(x = reorder_within(x=term, by=beta, within=topic_name),
                              y = beta,
                              fill = factor(topic_name))) +
                   geom_col(show.legend = F) +
                   facet_wrap(~ topic_name, scales = "free", ncol = 4) +
                   coord_flip() +
                   scale_x_reordered() +
                   scale_y_continuous(n.breaks = 4, # 축 눈금을 4개 내외로 정하기
                   labels = number_format(accuracy = .01)) + # 눈금 소수점 첫째 자리에서 반올림
                   labs(x = NULL) +
                   theme(text = element_text(family = "nanumgothic"))
```

Q8. LDA 모델의 beta를 이용해 각 토픽에 등장할 확률이 높은 상위 10개 단어를 추출한 다음 토픽 별 주요 단어를 나타낸 막대 그래프를 만드시오.

- R 결과




| | topic | term | beta | topic_name |
|---|---|---|---|---|
| 1 | 1 | 회당 | 0.007623774 | Topic 1 |
| 2 | 1 | 개성 | 0.007459998 | Topic 1 |
| 3 | 1 | 대학교 | 0.007050558 | Topic 1 |
| 4 | 1 | 자율 | 0.006231678 | Topic 1 |
| 5 | 1 | 이행 | 0.005904126 | Topic 1 |
| 6 | 1 | 남북간 | 0.005576573 | Topic 1 |
| 7 | 1 | 통일 | 0.005412797 | Topic 1 |
| 8 | 1 | 체제 | 0.005330909 | Topic 1 |
| 9 | 1 | 선언 | 0.005249021 | Topic 1 |
| 10 | 1 | 아이들 | 0.005003357 | Topic 1 |
| 11 | 2 | 지역주의 | 0.005366010 | Topic 2 |
| 12 | 2 | 야당 | 0.004996365 | Topic 2 |
| 13 | 2 | 당원 | 0.004626720 | Topic 2 |
| 14 | 2 | 대선 | 0.004565113 | Topic 2 |
| 15 | 2 | 총선 | 0.004565113 | Topic 2 |
| 16 | 2 | 정치권 | 0.004195468 | Topic 2 |
| 17 | 2 | 지역구도 | 0.004010646 | Topic 2 |
| 18 | 2 | 셀득 | 0.003887431 | Topic 2 |
| 19 | 2 | 권한 | 0.003825823 | Topic 2 |
| 20 | 2 | 이야기 | 0.003825823 | Topic 2 |
| 21 | 3 | 저력 | 0.007856391 | Topic 3 |
| 22 | 3 | 지도력 | 0.007321336 | Topic 3 |
| 23 | 3 | 동반자 | 0.006875457 | Topic 3 |
| 24 | 3 | 세듀니 | 0.006518754 | Topic 3 |

Showing 1 to 24 of 90 entries, 4 total columns

# Q9. LDA 모델의 gamma를 이용해 연설문 원문을 확률이 가장 높은 토픽으로 분류하시오.

- R 프로그램

```r
#Q9.
# gamma 추출
doc_topic <- tidy(lda_model, matrix = "gamma") %>% # 문서(id)가 토픽에 들어갈 확률을 포함한 df
             mutate(topic_name = paste("Topic", topic)) %>% print() # 토픽 이름 변수 설정
# gamma 확인
doc_topic %>% count(topic)

# 문서 별로 확률이 가장 높은 토픽 추출
doc_class <- doc_topic %>%
             group_by(document) %>%
             slice_max(gamma, n = 1) %>% print() # top_n(n=1, wt=gamma)과 동일함

# gamma가 동점이 발생한 경우 확인
doc_topic %>% group_by(document) %>%
              top_n(n=1, wt=gamma) %>%
              count(document) %>%
              filter(n > 1) %>% print()

# 데이터셋을 결합하기 위해 기준 변수 타입을 integer로 통일
doc_class$document <- as.integer(doc_class$document)

# 원문에 토픽 번호 부여
speech_content_topic <- raw_speech %>%
                        left_join(doc_class, by = c("id" = "document")) %>% print()
# 결합 확인
speech_content_topic %>% select(id, topic)
```

# Q9. LDA 모델의 gamma를 이용해 연설문 원문을 확률이 가장 높은 토픽으로 분류하시오.

- R 결과

| event | source | paragraph | content | id | topic | gamma | topic_name |
|---|---|---|---|---|---|---|---|
| 취임사 | 노무현대통령연설문집 제1권 2일 | 1 | 주한외교사절 여러분, 그리고 멀리서 오신 외빈 여러분, 전두... | 1 | 9 | 0.1748493 | Topic 9 |
| 취임사 | 노무현대통령연설문집 제1권 2일 | 1 | 존경하는 국민 여러분. 오늘 저는 대한민국의 제16대 대통령... | 2 | 9 | 0.2350304 | Topic 9 |
| 기념사 | 노무현대통령연설문집 제1권 2일 | 1 | 존경하는 본 바이제커 전 독일 대통령, 나카소네 야스히로 전... | 3 | 8 | 0.1784512 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 2일 | 1 | 43년 전 오늘, 대구는 자유당 정권의 독재에 맞서 분연히 일... | 4 | 6 | 0.2622739 | Topic 6 |
| 기념사 | 노무현대통령연설문집 제1권 2일 | 1 | 친애하는 학군 제41기 신임장교 여러분, 그리고 학부모와 내... | 5 | 8 | 0.5854701 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 3일 | 1 | 존경하는 국민 여러분, 오늘 여든 네 번째 3.1절을 맞이 나라... | 6 | 6 | 0.4183874 | Topic 6 |
| 기타 | 노무현대통령연설문집 제1권 3일 | 1 | ■ 인사말씀 안녕하십니까, 반갑습니다. 오늘 토론회에 참석... | 7 | 4 | 0.4304450 | Topic 4 |
| 기념사 | 노무현대통령연설문집 제1권 3일 | 1 | 친애하는 육군사관학교 졸업생 여러분, 학부모와 사관생도, ... | 8 | 8 | 0.6036325 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 3일 | 1 | 친애하는 해군사관학교 졸업생 여러분, 자리를 함께 하신 학... | 9 | 8 | 0.6201814 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 3일 | 1 | 존경하는 마산 시민 여러분! 오늘은 자유당 정권의 독재에 항... | 10 | 6 | 0.3212560 | Topic 6 |
| 기념사 | 노무현대통령연설문집 제1권 3일 | 1 | 친애하는 공군사관학교 졸업생 여러분, 사관생도와 공군장... | 11 | 8 | 0.6118264 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 3일 | 1 | 존경하는 상공인 여러분, 이 자리에 참석하신 내외 귀빈 여러... | 12 | 5 | 0.2382479 | Topic 5 |
| 기념사 | 노무현대통령연설문집 제1권 3일 | 1 | 경찰대학 19기 졸업생 여러분, 학부모와 내외 귀빈 여러분, ... | 13 | 8 | 0.4456019 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 3일 | 1 | 존경하는 제주도민 여러분, 안녕하십니까, 제주 국제컨벤션... | 14 | 7 | 0.3172840 | Topic 7 |
| 기념사 | 노무현대통령연설문집 제1권 3일 | 1 | 친애하는 육군3사관학교 졸업생 여러분, 학부모와 사관생도,... | 15 | 8 | 0.6144089 | Topic 8 |
| 국회연설 | 노무현대통령연설문집 제1권 4일 | 1 | 존경하는 국회의장, 그리고 국회의원 여러분, 이 자리를 지켜... | 16 | 2 | 0.4148969 | Topic 2 |
| 기념사 | 노무현대통령연설문집 제1권 4일 | 1 | 향토예비군 창설 서른 다섯 돌을 진심으로 축하합니다. 지역... | 17 | 8 | 0.3864734 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 4일 | 1 | 제1회 '동북아경제포럼'의 개막을 축하합니다. 해외에서 오... | 18 | 9 | 0.2183007 | Topic 9 |
| 기념사 | 노무현대통령연설문집 제1권 4일 | 1 | 친애하는 해군장병 여러분, 그리고 현대중공업 임직원과 내... | 19 | 8 | 0.5772080 | Topic 8 |

Showing 1 to 19 of 802 entries, 12 total columns

# Q10. 토픽 별 문서 수를 출력하시오.

- R 프로그램

```
#Q10.
# 토픽 별 문서 수 확인
speech_content_topic <- news_comment_topic %>%
                        na.omit()
speech_content_topic %>% count(topic)
```

- R 결과

| | topic | n |
|---|---|---|
| **1** | 1 | 3 |
| **2** | 2 | 5 |
| **3** | 3 | 15 |
| **4** | 4 | 6 |
| **5** | 5 | 21 |
| **6** | 6 | 19 |
| **7** | 7 | 42 |
| **8** | 8 | 30 |
| **9** | 9 | 37 |

Q11. 문서가 가장 많은 토픽의 연설문을 gamma가 높은 순으로 출력하고 내용이 비슷한지 살펴보시오.

- R 프로그램

```
#Q11.
# gamma가 높은 주요 문서가 먼저 출력되도록 정렬
content_topic <- speech_content_topic %>%
                 arrange(-gamma)
content_topic %>% select(topic, gamma, content)
```

# Q11. 문서가 가장 많은 토픽의 연설문을 gamma가 높은 순으로 출력하고 내용이 비슷한지 살펴보시오.

- R 결과

| event | source | paragraph | content | id | topic | gamma | topic_name |
|---|---|---|---|---|---|---|---|
| 성명/담화문 | 노무현대통령연설문집 제1권 11월 | 1 | 존경하는 전국의 농업인 여러분, 그리고 농업 지도자 여러분,... | 127 | 5 | 0.6352555 | Topic 5 |
| 성명/담화문 | 노무현대통령연설문집 제1권 11월 | 1 | 존경하는 박관용 국회의장, 그리고 국회의원 여러분, 의정 활... | 134 | 5 | 0.6304811 | Topic 5 |
| 국회연설 | 노무현대통령연설문집 제1권 6월 | 1 | 존경하는 국회의장, 그리고 국회의원 여러분! 2003년도 제1... | 55 | 5 | 0.6207133 | Topic 5 |
| 기념사 | 노무현대통령연설문집 제1권 3월 | 1 | 진애하는 해군사관학교 졸업생 여러분, 자리를 함께 하신 학... | 9 | 8 | 0.6201814 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 3월 | 1 | 진애하는 육군3사관학교 졸업생 여러분, 학부모와 사관생도,... | 15 | 8 | 0.6144089 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 3월 | 1 | 진애하는 공군사관학교 졸업생 여러분, 사관생도와 공군장... | 11 | 8 | 0.6118264 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 3월 | 1 | 진애하는 육군사관학교 졸업생 여러분, 학부모와 사관생도, ... | 8 | 8 | 0.6036325 | Topic 8 |
| 기타 | 노무현대통령연설문집 제1권 10월 | 1 | 존경하는 루디 패석 ASEAN 민간자문위원회 위원장 타리 아... | 104 | 9 | 0.5922068 | Topic 9 |
| 국회연설 | 노무현대통령연설문집 제1권 7월 | 1 | 존경하는 박관용 국회의장, 그리고 국회의원 여러분. 저는 제... | 64 | 5 | 0.5895062 | Topic 5 |
| 기념사 | 노무현대통령연설문집 제1권 2월 | 1 | 진애하는 학군 제41기 신임장교 여러분, 그리고 학부모와 내... | 5 | 8 | 0.5854701 | Topic 8 |
| 기념사 | 노무현대통령연설문집 제1권 8월 | 1 | 존경하는 조석래 태평양경제협의회(PBEC) 회장, 그리고 아, ... | 84 | 9 | 0.5822731 | Topic 9 |
| 기념사 | 노무현대통령연설문집 제1권 4월 | 1 | 진애하는 해군장병 여러분, 그리고 현대중공업 임직원과 내... | 19 | 8 | 0.5772080 | Topic 8 |
| 기타 | 노무현대통령연설문집 제1권 10월 | 1 | 존경하는 국민여러분 저는 인도네시아 발리에서 열린 아세... | 106 | 9 | 0.5707071 | Topic 9 |
| 기념사 | 노무현대통령연설문집 제1권 10월 | 1 | 진애하는 국군장병 여러분, 그리고 내외귀빈 여러분, 오늘은 ... | 100 | 8 | 0.5591985 | Topic 8 |
| 국회연설 | 노무현대통령연설문집 제1권 10월 | 1 | 존경하는 국회의장, 그리고 국회의원 여러분! 2003년도 제2... | 105 | 5 | 0.5514212 | Topic 5 |
| 성명/담화문 | 노무현대통령연설문집 제1권 7월 | 1 | 존경하는 국민 여러분, 저는 최근 대선자금에 관한 사회적 공... | 72 | 2 | 0.5344982 | Topic 2 |
| 기념사 | 노무현대통령연설문집 제1권 11월 | 1 | 국방일보 '주먹의 내무반'이 100회를 맞게 된 것을 진심으로 ... | 123 | 8 | 0.5336833 | Topic 8 |
| 성명/담화문 | 노무현대통령연설문집 제1권 7월 | 1 | 존경하는 한 중 경제인 여러분, 먼저, 저와 우리 일행을 이처... | 69 | 9 | 0.5286369 | Topic 9 |
| 기타 | 노무현대통령연설문집 제1권 12월 | 1 | 존경하는 내외귀빈 여러분, '동아시아 포럼'의 창립을 진심으... | 151 | 9 | 0.5284722 | Topic 9 |

Showing 2 to 20 of 178 entries, 12 total columns

- gamma 값이 높은 기준으로, 같은 토픽끼리 비교한 결과 내용이 유사한 것을 알 수 있음.