텍스트 마이닝(4)

숙명여자대학교 경영학부 오중산

오즈비: 상대적으로 중요한 단어 비교하기

- 단어 빈도 비교하기의 한계
  - 어떤 텍스트에서든 많이 사용되는 범용적 단어의 경우 비교의 의미가 없음
    - 예: "우리", "사회", "경제", "일자리"
  - 특정 텍스트에서는 많이 사용되지만, 다른 텍스트에서는 덜 사용되는 '상대적 빈도가 높은 단어'가 무엇인지 파악하는 것이 중요함

오즈비: 상대적으로 중요한 단어 비교하기

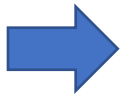
- Long form 형태 데이터를 wide form으로 바꾸기
  - ◆ Long form 형태 데이터의 한계
    - ❖ 같은 단어가 범주별로 다른 행을 구성하여 빈도 비교가 어렵고, 연산하기도 불편함
    - ❖ df_long <- frequency %>% group_by(president) %>% slice_max(n, n = 10) %>% filter(word %in% c("국민", "우리", "정치", "행복"))

```
# A tibble: 6 x 3
# Groups:   president [2]
  president word      n
  <chr>     <chr> <int>
1 moon      국민     21
2 moon      우리     17
3 moon      정치     12
4 park      국민     72
5 park      행복     23
6 park      우리     10
```

오즈비: 상대적으로 중요한 단어 비교하기

- Long form 형태 데이터를 wide form으로 바꾸기
  - ◆ tidyr 패키지의 pivot_wider 함수를 이용하기
    - ❖ names_from: 변수명을 가져올 변수
    - ❖ values_from: 변수에 채워 넣을 값이 있는 변수
    - ❖ df_wide <- df_long %>% pivot_wider(names_from = president, values_from = n, values_fill = list(n = 0))

```
# A tibble: 4 x 3
  word   moon  park
  <chr> <int> <int>
1 국민     21    72
2 우리     17    10
3 정치     12    NA
4 행복     NA    23
```



```
# A tibble: 4 x 3
  word   moon  park
  <chr> <int> <int>
1 국민     21    72
2 우리     17    10
3 정치     12     0
4 행복      0    23
```

오즈비: 상대적으로 중요한 단어 비교하기

- 연설문 단어 빈도 데이터 프레임(frequency)을 wide form으로 바꾸기
  - ◆ 두 대통령의 연설문 단어 빈도를 저장한 frequency를 wide form으로 변경
    - ❖ frequency_wide <- frequency %>% pivot_wider(names_from = president, values_from = n, values_fill = list(n = 0))

```
# A tibble: 955 x 3
   word       moon  park
   <chr>     <int> <int>
 1 가동          1     0
 2 가사          1     0
 3 가슴          2     0
 4 가족          1     1
 5 가족구조      1     0
 6 가지          4     0
 7 가치          3     1
 8 각종          1     0
 9 감당          1     0
10 강력          3     0
# ... with 945 more rows
```

오즈비: 상대적으로 중요한 단어 비교하기

- 오즈비(odds ratio) 구하기
  - 오즈비란?
    - 어떤 사건의 A조건에서 발생 확률이 B조건에서 발생할 확률에 비해 얼마나 더 큰지 나타낸 값
    - 단어가 두 텍스트 중 어디 등장할 확률이 높은지, 즉 단어의 상대적인 중요도를 알 수 있음
  - 연설문별 단어 비중 구하기
    - 연설문별로 '각 단어의 빈도'를 '모든 단어 빈도의 합'으로 나눔
    - 단어 빈도가 0이면 오즈비 구할 때 문제가 발생해서 분모/분자에 각각 1을 더함
    - frequency_wide <- frequency_wide %>% mutate(ratio_moon = ((moon + 1)/(sum(moon + 1))), ratio_park = ((park + 1)/(sum(park + 1))))

오즈비: 상대적으로 중요한 단어 비교하기

- 오즈비(odds ratio) 구하기
  - ◆ 오즈비 변수 추가하기
    - ❖ 한 텍스트의 단어 비중을 다른 텍스트의 단어 비중으로 나눔
    - ❖ frequency_wide <- frequency_wide %>% mutate(odds_ratio = ratio_moon/ratio_park)

```
# A tibble: 955 x 6
   word      moon  park ratio_moon ratio_park odds_ratio
   <chr>    <int> <int>      <dbl>      <dbl>      <dbl>
 1 가동         1     0   0.000873   0.000552      1.58
 2 가사         1     0   0.000873   0.000552      1.58
 3 가슴         2     0   0.00131    0.000552      2.37
 4 가족         1     1   0.000873   0.00110       0.791
 5 가족구조     1     0   0.000873   0.000552      1.58
 6 가지         4     0   0.00218    0.000552      3.96
 7 가치         3     1   0.00175    0.00110       1.58
 8 각종         1     0   0.000873   0.000552      1.58
 9 감당         1     0   0.000873   0.000552      1.58
10 강력         3     0   0.00175    0.000552      3.17
# ... with 945 more rows
```

오즈비: 상대적으로 중요한 단어 비교하기

- 오즈비 해석
  - ◆ 오즈비가 1보다 크거나 작을 때의 의미
    - ❖ 1보다 크면, 박 전 대통령 연설문에 비해 문대통령 연설문에서 해당 단어가 더 많이 사용됨
      - ✓ frequency_wide %>% arrange(-odds_ratio)
    - ❖ 1보다 작으면, 문대통령 연설문에 비해 박 전 대통령 연설문에서 해당 단어가 더 많이 사용됨
      - ✓ frequency_wide %>% arrange(odds_ratio)
    - ❖ 두 연설문에서 비중이 동일하면 오즈비는 1이됨
      - ✓ frequency_wide %>% arrange(abs(1 - odds_ratio))

오즈비: 상대적으로 중요한 단어 비교하기

- 오즈비가 가장 높은 10개 단어와 가장 낮은 10개 단어 추출하여 top10 만들기
  - ◆ 전자는 문대통령 연설문에서 상대적으로 비중이 더 높음
  - ◆ 후자는 박 전 대통령 연설문에서 상대적으로 비중이 더 높음
  - ◆ 값이 작은 순서대로 순위를 구하는 rank 함수 이용
    - ❖ top10 <- frequency_wide %>% filter(rank(odds_ratio) <= 10 | rank(-odds_ratio) <= 10) %>% arrange(-odds_ratio)

```
# A tibble: 20 x 6
   word      moon  park ratio_moon ratio_park odds_ratio
   <chr>    <int> <int>      <dbl>      <dbl>      <dbl>
 1 복지국가     8     0    0.00393   0.000552       7.12
 2 세상         6     0    0.00306   0.000552       5.54
 3 여성         6     0    0.00306   0.000552       5.54
 4 정의         6     0    0.00306   0.000552       5.54
 5 강자         5     0    0.00262   0.000552       4.75
 6 공평         5     0    0.00262   0.000552       4.75
 7 대통령의     5     0    0.00262   0.000552       4.75
 8 보통         5     0    0.00262   0.000552       4.75
 9 상생         5     0    0.00262   0.000552       4.75
10 지방         5     0    0.00262   0.000552       4.75
11 과제         0     4   0.000436    0.00276      0.158
12 국정운영     0     4   0.000436    0.00276      0.158
13 시작         0     4   0.000436    0.00276      0.158
14 지식         0     4   0.000436    0.00276      0.158
15 행복         3    23    0.00175     0.0132      0.132
16 실천         0     5   0.000436    0.00331      0.132
17 정보         0     5   0.000436    0.00331      0.132
18 투명         0     5   0.000436    0.00331      0.132
19 여러분       2    20    0.00131     0.0116      0.113
20 박근혜       0     8   0.000436    0.00496     0.0879
```

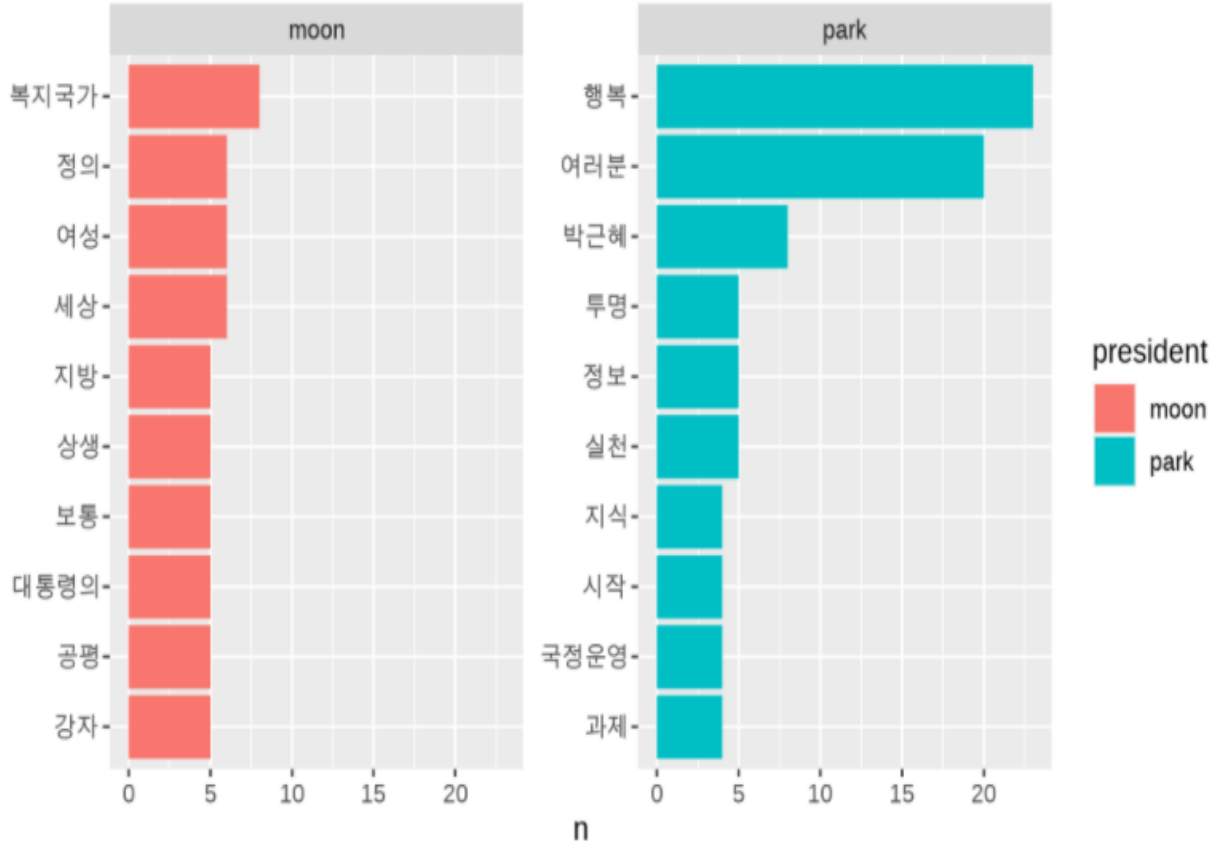
오즈비: 상대적으로 중요한 단어 비교하기

- 막대 그래프 그리기
  - ◆ 막대 그래프를 그리기 위한 새로운 변수(president와 n) 만들기
    - ❖ 해당 단어가 어느 대통령 연설문에서 비롯된 것인지, 그리고 그 빈도가 얼마인지 알기 위해 두 개의 새로운 변수 형성
    - ❖ top10 <- top10 %>% mutate(president = ifelse(odds_ratio > 1, "moon", "park"), n = ifelse(odds_ratio > 1, moon, park))

```
# A tibble: 20 x 8
   word      moon  park ratio_moon ratio_park odds_ratio president     n
   <chr>    <int> <int>      <dbl>      <dbl>      <dbl> <chr>     <int>
 1 강자         5     0    0.00262   0.000552       4.75  moon          5
 2 공평         5     0    0.00262   0.000552       4.75  moon          5
 3 대통령의     5     0    0.00262   0.000552       4.75  moon          5
```

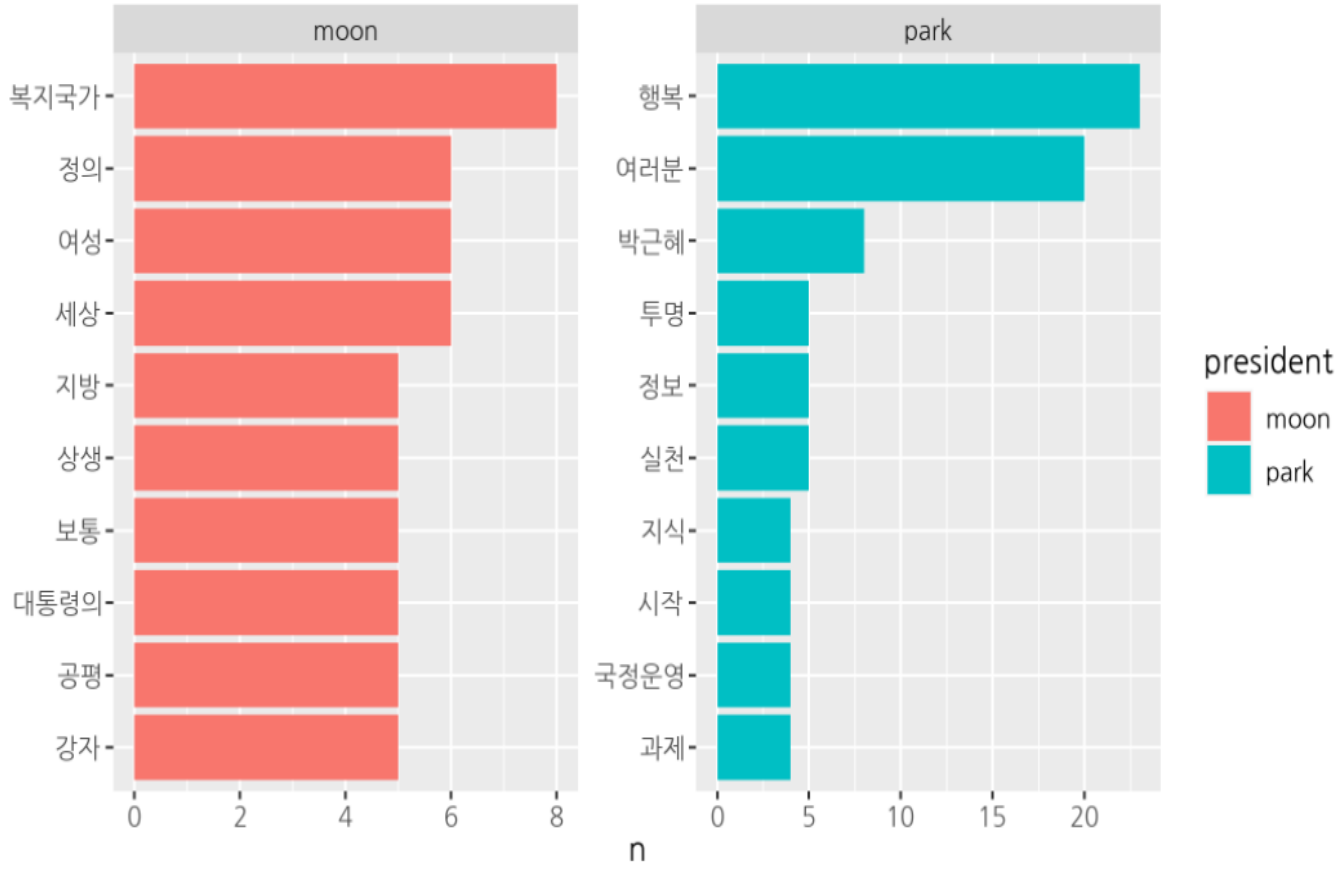
오즈비: 상대적으로 중요한 단어 비교하기

- 막대 그래프 그리기
  - ◆ 두 대통령 연설문에서 상대적으로 많이 사용된 단어 빈도 비교
    - ❖ ggplot(top10, aes(x = reorder_within(word, n, president), n, fill = president)) + geom_bar(stat = "identity") + coord_flip() + facet_wrap(~ president, scales = "free_y") + scale_x_reordered() + labs(x = NULL)
    - ❖ 주의! scale_x_reordered 함수를 사용하려면 tidytext 패키지를 불러와야 함



오즈비: 상대적으로 중요한 단어 비교하기

- 막대 그래프 그리기
  - ◆ 그래프별로 축 설정하기
    - ❖ ggplot(top10, aes(x = reorder_within(word, n, president), n, fill = president)) + geom_bar(stat = "identity") + coord_flip() + facet_wrap(~ president, scales = "free") + scale_x_reordered() + labs(x = NULL)
    - ❖ 막대 길이가 같아도 빈도가 다름에 주의해야 함



오즈비: 상대적으로 중요한 단어 비교하기

- 주요 단어가 사용된 문장 살펴보기
  - ◆ 문장 단위로 토큰화하기
    - ❖ speeches_sentence <- bind_speeches %>% as_tibble() %>% unnest_tokens(input = value, output = sentence, token = "sentences")
  - ◆ 주요 단어가 사용된 문장 추출하기
    - ❖ speeches_sentence %>% filter(president == "moon" & str_detect(sentence, "복지국가"))
    - ❖ speeches_sentence %>% filter(president == "park" & str_detect(sentence, "행복"))

오즈비: 상대적으로 중요한 단어 비교하기

- 중요도가 비슷한 단어 살펴보기
  - 오즈비가 1에 가까운 보편적인 단어
    - frequency_wide %>% arrange(abs(1 - odds_ratio)) %>% head(10)
- 중요도가 비슷하고 빈도가 높은 단어 살펴보기
  - 각 연설문에서 빈도수가 5회 이상이면서 중요도가 비슷한 단어 추출
  - 두 연설문에서 모두 강조한 단어 확인
    - frequency_wide %>% filter(moon >= 5 & park >= 5) %>% arrange(abs(1 - odds_ratio)) %>% head(10)

```
# A tibble: 10 x 6
   word      moon  park ratio_moon ratio_park odds_ratio
   <chr>    <int> <int>      <dbl>      <dbl>      <dbl>
 1 사회        14     9    0.00655    0.00552      1.19
 2 사람         9     9    0.00436    0.00552      0.791
 3 경제        15    15    0.00698    0.00883      0.791
 4 지원         5     5    0.00262    0.00331      0.791
 5 우리        17    10    0.00786    0.00607      1.29
 6 불안         7     8    0.00349    0.00496      0.703
 7 산업         9     5    0.00436    0.00331      1.32
 8 대한민국    11     6    0.00524    0.00386      1.36
 9 국가         7    10    0.00349    0.00607      0.576
10 교육         6     9    0.00306    0.00552      0.554
```