동국대학교
dongguk university

실습 8강
토픽모델링

실습 데이터

- Topicmodels에 내장된 "AssociatedPress“ 데이터 셋
- 미국의 2246 개 뉴스 기사 모음

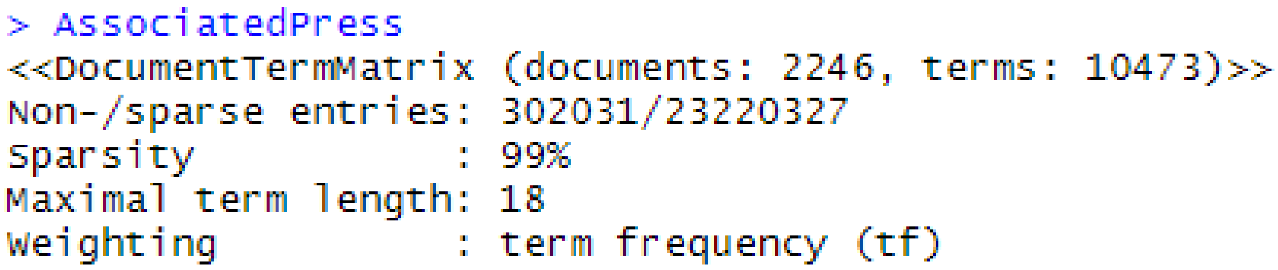
- DTM 형태로 되어 있음

데이터 준비

```
# 라이브러리 로딩
library(topicmodels)
library(tidytext)
library(tidyr)
library(ggplot2)
library(dplyr)

# DTM 예제 데이터 로딩
data("AssociatedPress")

# 예제 데이터 확인
AssociatedPress
```

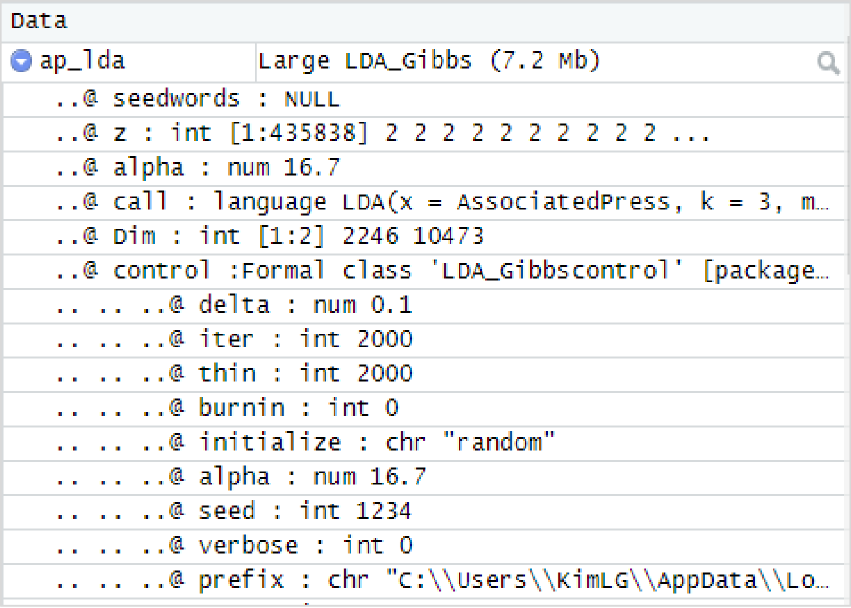


```
> AssociatedPress
<<DocumentTermMatrix (documents: 2246, terms: 10473)>>
Non-/sparse entries: 302031/23220327
Sparsity           : 99%
Maximal term length: 18
Weighting          : term frequency (tf)
```

LDA 모델링

```
# k : 토픽 수
# method :  "Gibbs" 선택
ap_lda <- lda(AssociatedPress,
              k = 3,
              method = "Gibbs",
              control = list(seed = 1234))
ap_lda
```

```
> ap_lda
A LDA_Gibbs topic model with 3 topics.
```

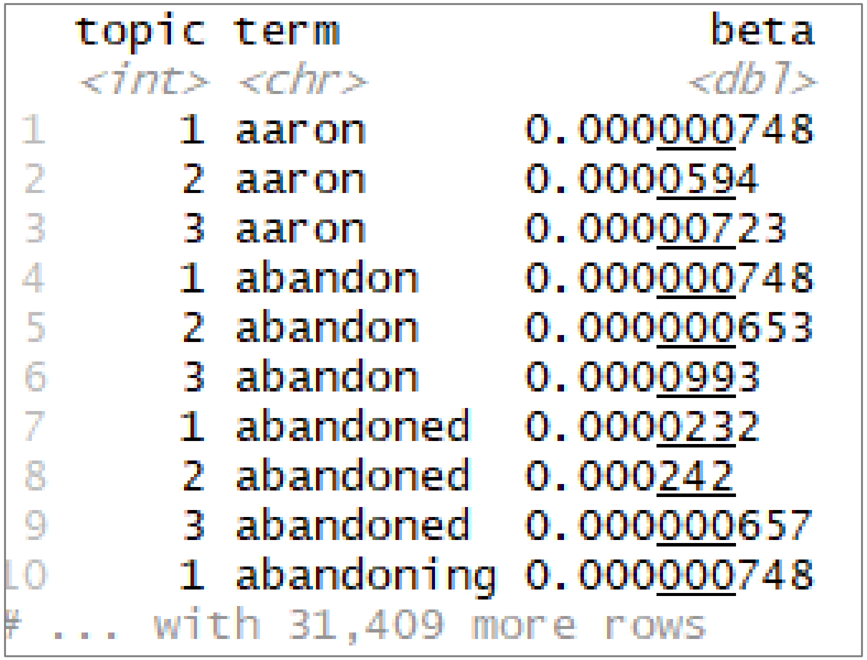


```
Data
ap_lda                Large LDA_Gibbs (7.2 Mb)
  ..@ seedwords : NULL
  ..@ z : int [1:435838] 2 2 2 2 2 2 2 2 2 2 ...
  ..@ alpha : num 16.7
  ..@ call : language LDA(x = AssociatedPress, k = 3, m…
  ..@ Dim : int [1:2] 2246 10473
  ..@ control :Formal class 'LDA_Gibbscontrol' [package…
  .. .. ..@ delta : num 0.1
  .. .. ..@ iter : int 2000
  .. .. ..@ thin : int 2000
  .. .. ..@ burnin : int 0
  .. .. ..@ initialize : chr "random"
  .. .. ..@ alpha : num 16.7
  .. .. ..@ seed : int 1234
  .. .. ..@ verbose : int 0
  .. .. ..@ prefix : chr "C:\\Users\\KimLG\\AppData\\Lo…
```

베타 탐색

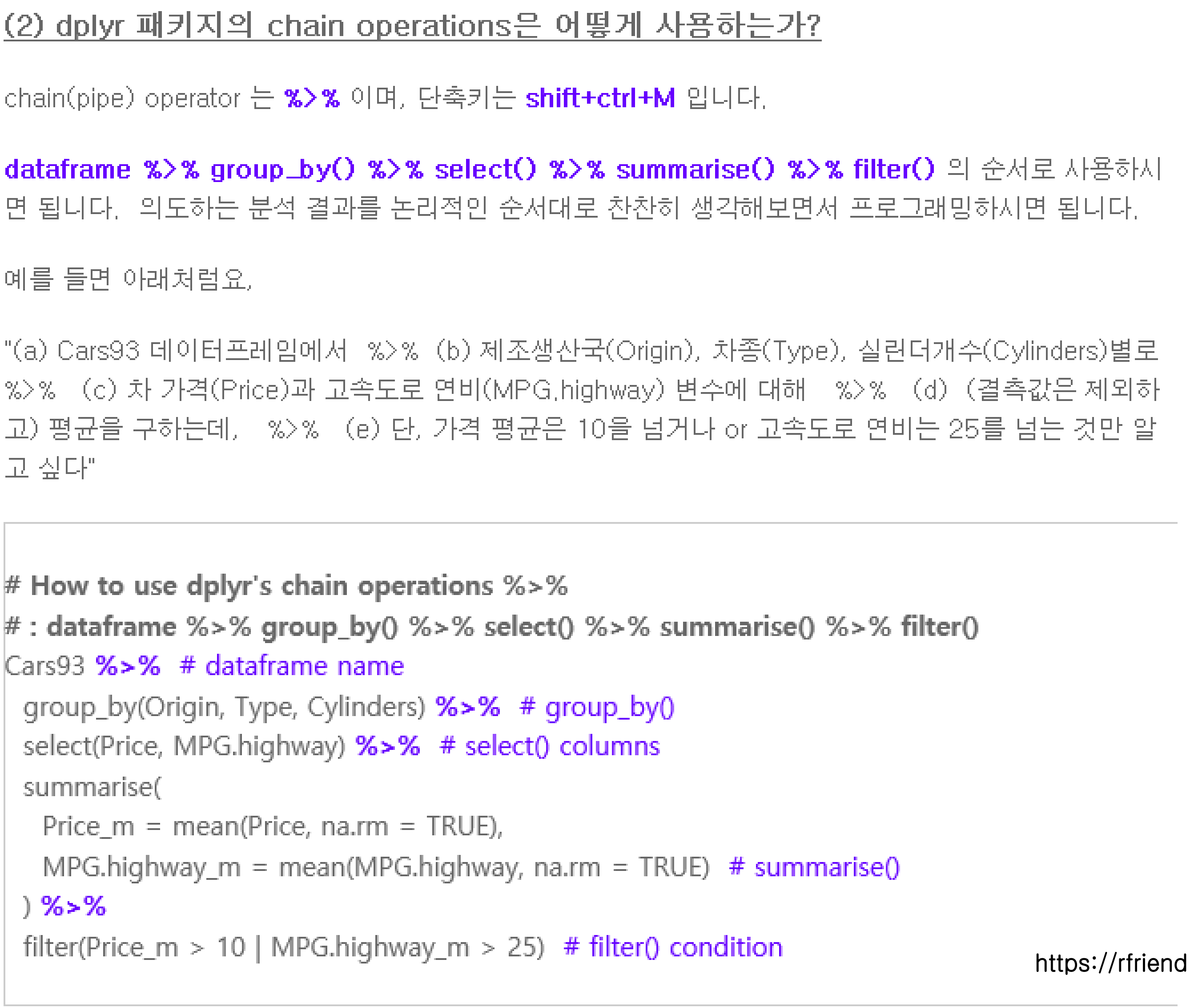
```
# tidy() : LDA 모형 결과 확인
# 구축된 모형으로부터 beta (토픽 별 단어 확률분포) 도출
# 도출 기준 : beta / gamma(문서 별 토픽 확률분포)

ap_topics <- tidy(ap_lda, matrix = "beta")
ap_topics
```



```
   topic term               beta
   <int> <chr>             <dbl>
 1     1 aaron       0.000000748
 2     2 aaron       0.0000594
 3     3 aaron       0.00000723
 4     1 abandon     0.000000748
 5     2 abandon     0.000000653
 6     3 abandon     0.0000993
 7     1 abandoned   0.0000232
 8     2 abandoned   0.000242
 9     3 abandoned   0.000000657
10     1 abandoning  0.000000748
# ... with 31,409 more rows
```

참고



(2) dplyr 패키지의 chain operations은 어떻게 사용하는가?

chain(pipe) operator 는 **%>%** 이며, 단축키는 **shift+ctrl+M** 입니다.

dataframe %>% group_by() %>% select() %>% summarise() %>% filter() 의 순서로 사용하시면 됩니다. 의도하는 분석 결과를 논리적인 순서대로 찬찬히 생각해보면서 프로그래밍하시면 됩니다.

예를 들면 아래처럼요.

"(a) Cars93 데이터프레임에서 %>% (b) 제조생산국(Origin), 차종(Type), 실린더개수(Cylinders)별로 %>% (c) 차 가격(Price)과 고속도로 연비(MPG.highway) 변수에 대해 %>% (d) (결측값은 제외하고) 평균을 구하는데, %>% (e) 단, 가격 평균은 10을 넘거나 or 고속도로 연비는 25를 넘는 것만 알고 싶다"

```
# How to use dplyr's chain operations %>%
# : dataframe %>% group_by() %>% select() %>% summarise() %>% filter()
Cars93 %>%  # dataframe name
  group_by(Origin, Type, Cylinders) %>%  # group_by()
  select(Price, MPG.highway) %>%  # select() columns
  summarise(
    Price_m = mean(Price, na.rm = TRUE),
    MPG.highway_m = mean(MPG.highway, na.rm = TRUE)  # summarise()
  ) %>%
  filter(Price_m > 10 | MPG.highway_m > 25)  # filter() condition
```

https://rfriend.tistory.com/236

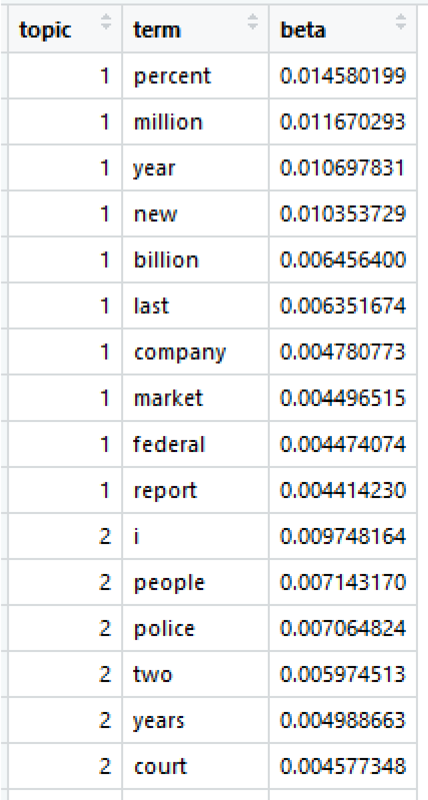
베타 탐색

- dplyr package
  - filter() : 지정 조건에 맞는 데이터(행) 추출
  - select() : 열 추출
  - mutate() : 열 추가
  - arrange() : 정렬

```
# 토픽 별 베타 정렬

ap_top_terms <- ap_topics %>%
  group_by(topic) %>%
  top_n(10, beta) %>%
  ungroup() %>%
  arrange(topic, -beta)

ap_top_terms
```



topic	term	beta
1	percent	0.014580199
1	million	0.011670293
1	year	0.010697831
1	new	0.010353729
1	billion	0.006456400
1	last	0.006351674
1	company	0.004780773
1	market	0.004496515
1	federal	0.004474074
1	report	0.004414230
2	i	0.009748164
2	people	0.007143170
2	police	0.007064824
2	two	0.005974513
2	years	0.004988663
2	court	0.004577348

베타 탐색

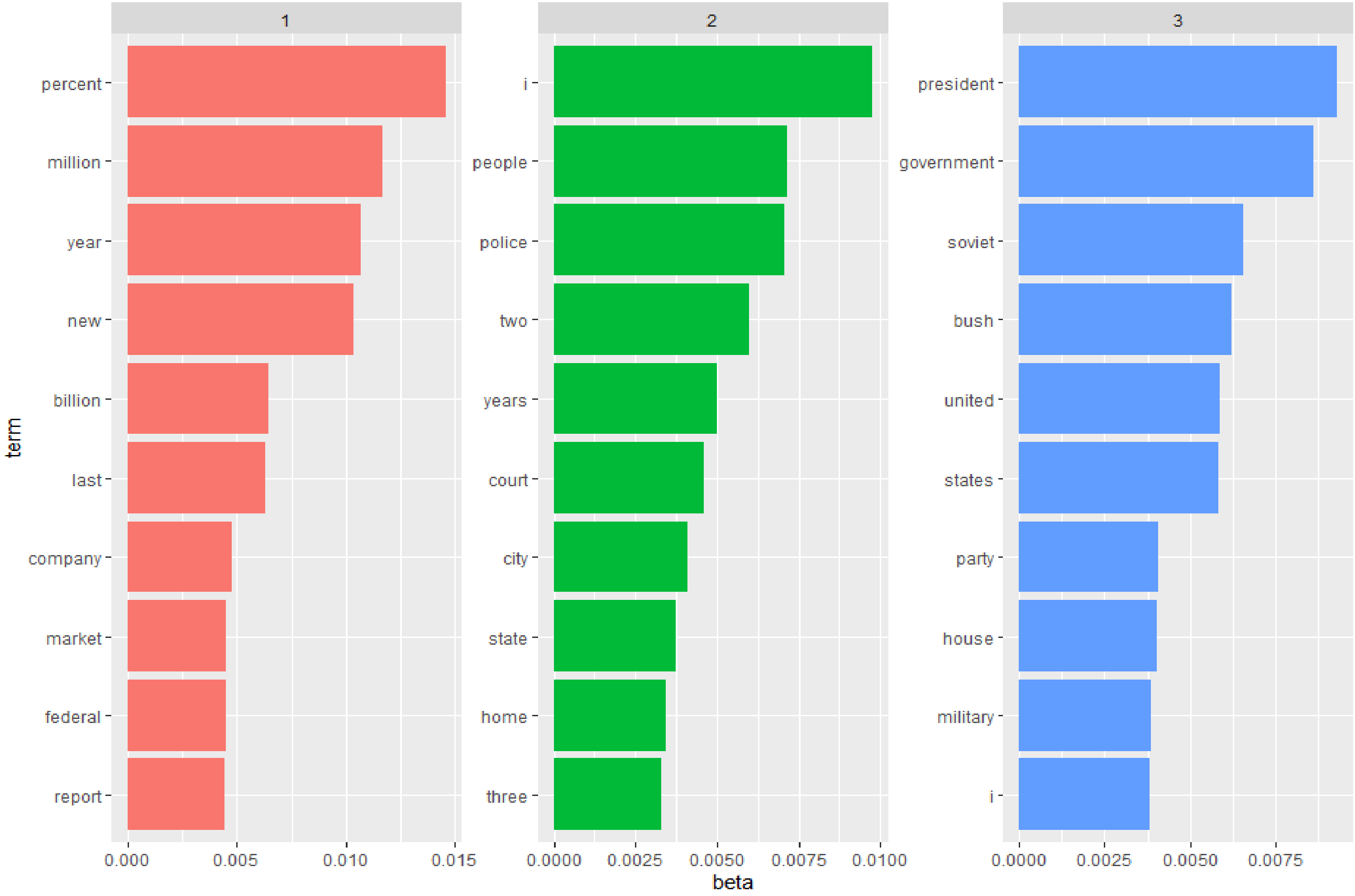
```
# ggplot() : 그림 입력을 "+" 이용하여 scale까지 표현
# geom_col : column 에 대한 정보
# facet_wrap() : 그래프 함수
# coord_flip() : 데이터 포인트 그리기
# scale_x_reordered : x축 재정렬

# ap_top_terms를 이용하여 ggplot 그리기

ap_top_terms %>%
  mutate(term = reorder_within(term, beta, topic)) %>%
  ggplot(aes(term, beta, fill = factor(topic))) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~ topic, scales = "free") +
  coord_flip() +
  scale_x_reordered()
```

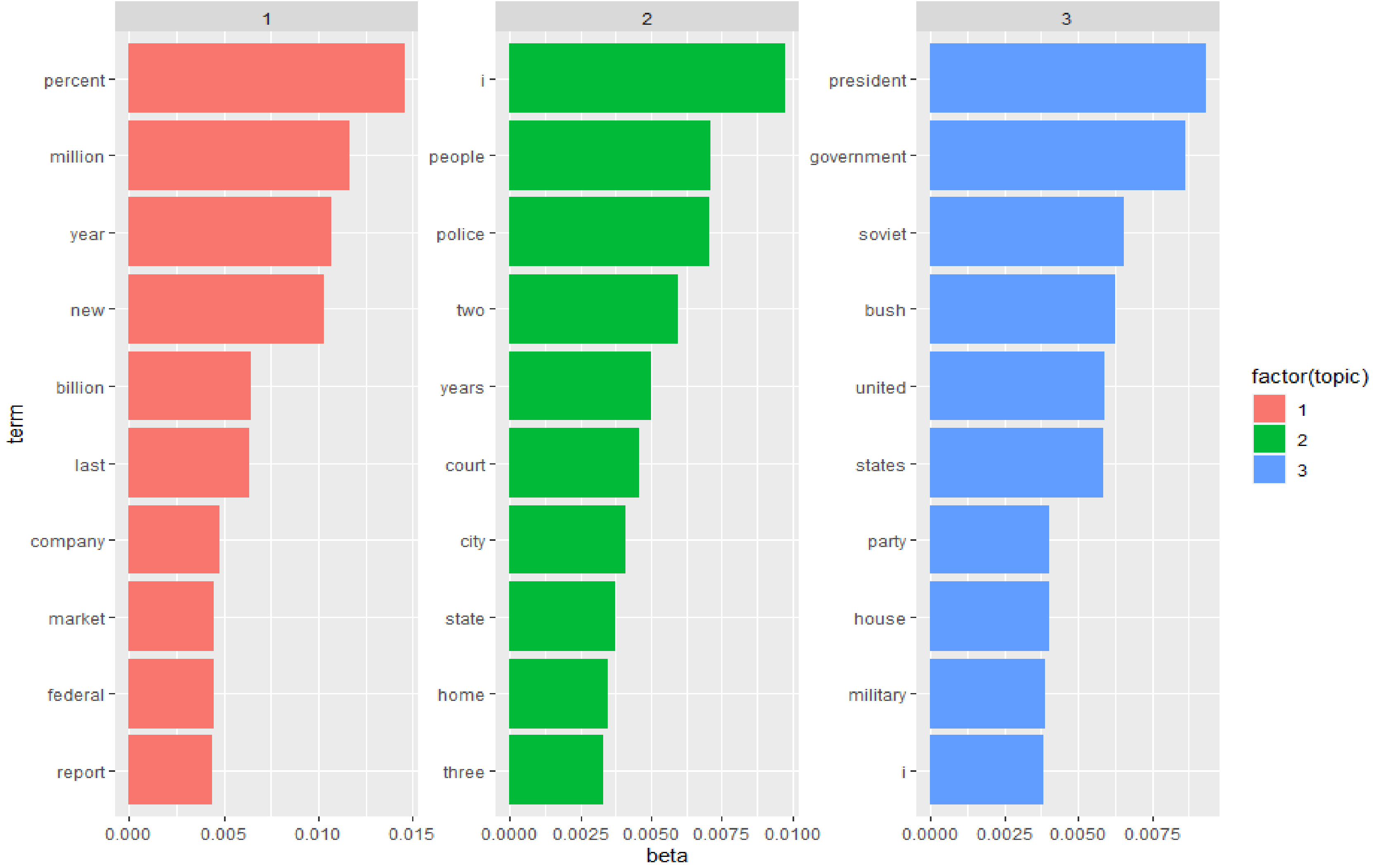
베타 탐색

- `geom_col(show.legend = FALSE)`



베타 탐색

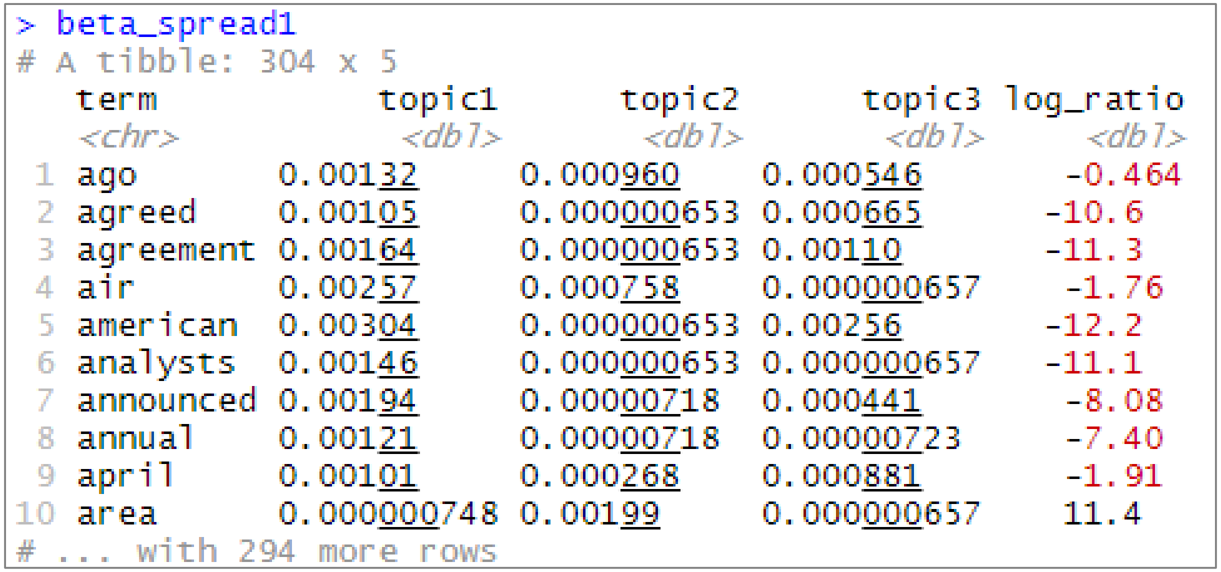
- geom_col(show.legend = TRUE)



토픽 별 쌍대비교

```
# spread() : value가 있는 다수 열을 선택

# beta_spread1 : topic 1과 topic2 비교
beta_spread1 <- ap_topics %>%
  mutate(topic = paste0("topic", topic)) %>%
  spread(topic, beta) %>%
  filter(topic1 > .001 | topic2 > .001) %>%
  mutate(log_ratio = log2(topic2 / topic1))
```

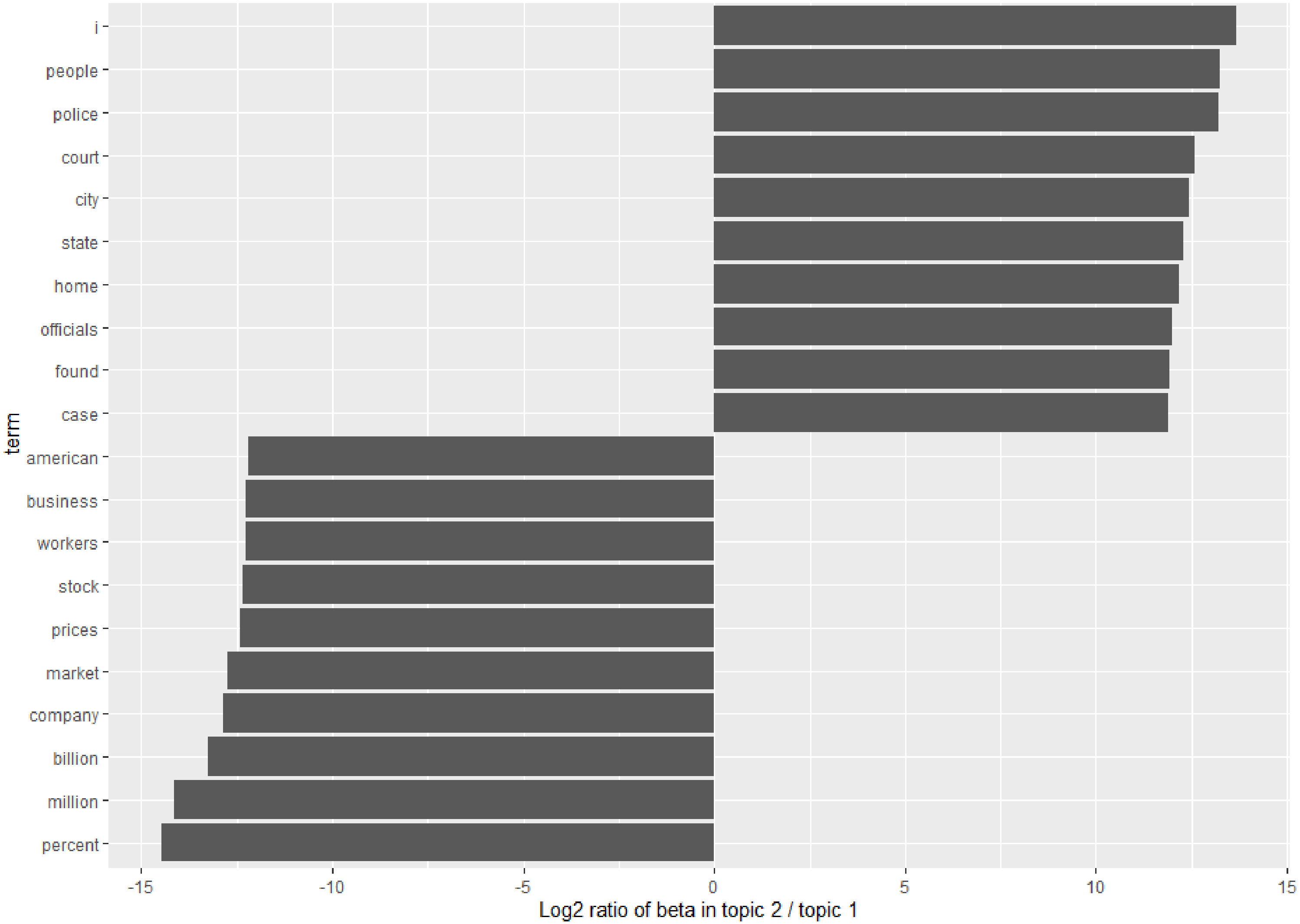


```
> beta_spread1
# A tibble: 304 x 5
   term           topic1      topic2      topic3 log_ratio
   <chr>           <dbl>       <dbl>       <dbl>     <dbl>
 1 ago       0.00132     0.000960    0.000546       -0.464
 2 agreed    0.00105     0.000000653 0.000665      -10.6
 3 agreement 0.00164     0.000000653 0.00110       -11.3
 4 air       0.00257     0.000758    0.000000657    -1.76
 5 american  0.00304     0.000000653 0.00256       -12.2
 6 analysts  0.00146     0.000000653 0.000000657   -11.1
 7 announced 0.00194     0.00000718  0.000441       -8.08
 8 annual    0.00121     0.00000718  0.00000723     -7.40
 9 april     0.00101     0.000268    0.000881       -1.91
10 area      0.000000748 0.00199     0.000000657    11.4
# ... with 294 more rows
```

토픽 별 쌍대비교

```
# ggplot을 이용하여 beta_spread1의 그래프 생성
beta_spread1 %>%
  group_by(direction = log_ratio > 0) %>%
  top_n(10, abs(log_ratio)) %>%
  ungroup() %>%
  mutate(term = reorder(term, log_ratio)) %>%
  ggplot(aes(term, log_ratio)) +
  geom_col() +
  labs(y = "Log2 ratio of beta in topic 2 / topic 1") +
  coord_flip()
```

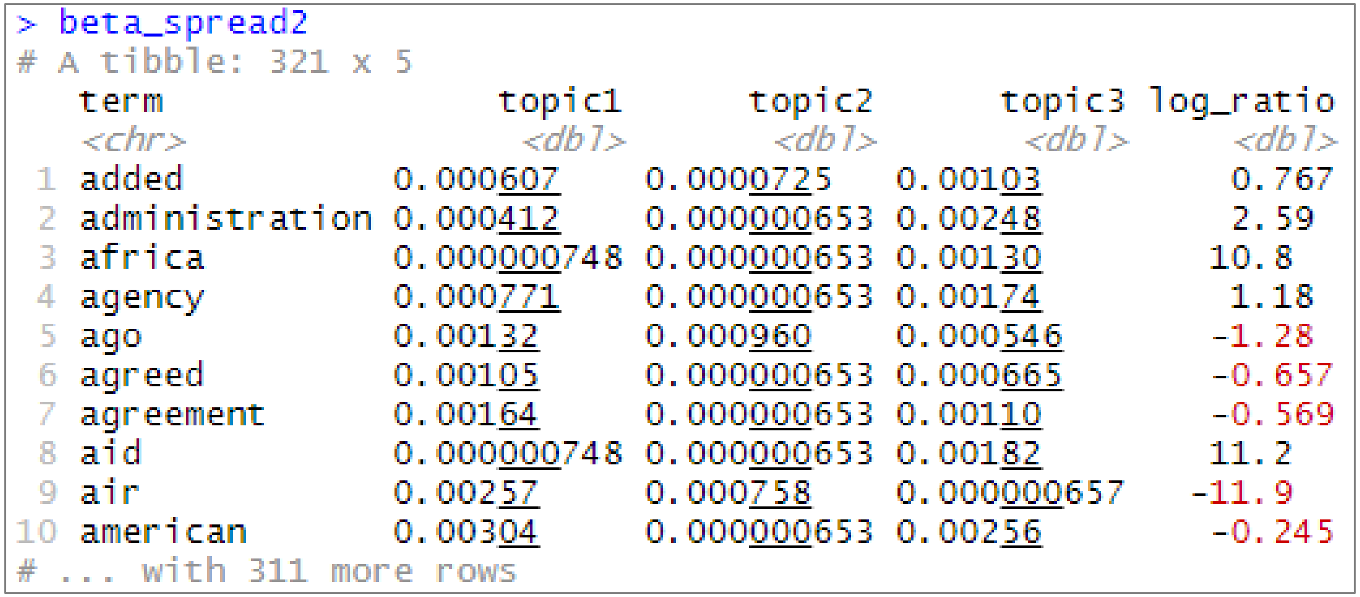
토픽 별 쌍대비교


i
people
police
court
city
state
home
officials
found
case
american
business
workers
stock
prices
market
company
billion
million
percent
term
-15
-10
-5
0
5
10
15
Log2 ratio of beta in topic 2 / topic 1

토픽 별 쌍대비교

```
# beta_spread2 : topic 1과 topic3 비교

beta_spread2 <- ap_topics %>%
  mutate(topic = paste0("topic", topic)) %>%
  spread(topic, beta) %>%
  filter(topic1 > .001 | topic3 > .001) %>%
  mutate(log_ratio = log2(topic3 / topic1))
```



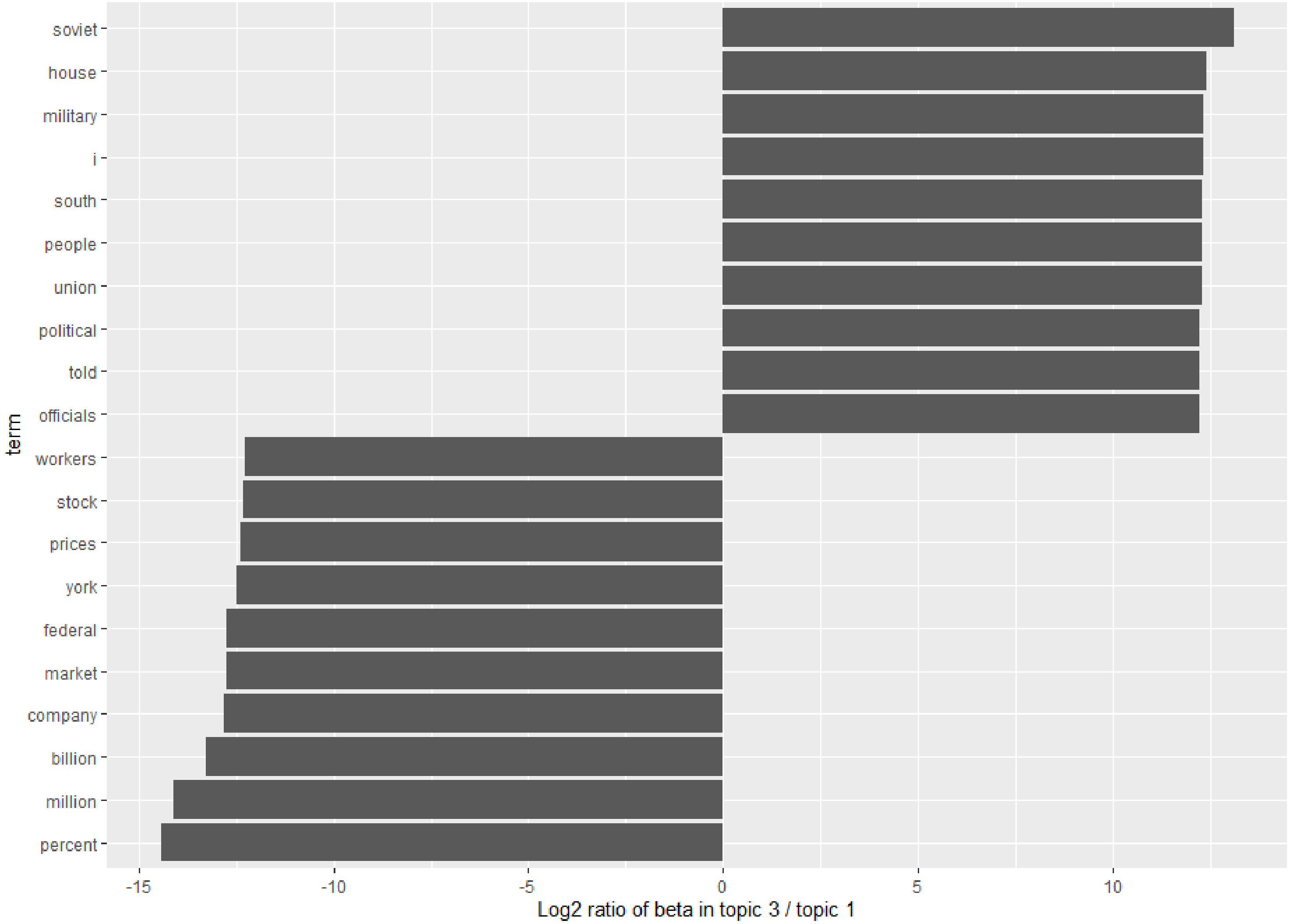
```
> beta_spread2
# A tibble: 321 x 5
   term               topic1      topic2      topic3 log_ratio
   <chr>               <dbl>       <dbl>       <dbl>     <dbl>
 1 added          0.000607    0.0000725   0.00103        0.767
 2 administration 0.000412    0.000000653 0.00248        2.59
 3 africa         0.000000748 0.000000653 0.00130       10.8
 4 agency         0.000771    0.000000653 0.00174        1.18
 5 ago            0.00132     0.000960    0.000546      -1.28
 6 agreed         0.00105     0.000000653 0.000665      -0.657
 7 agreement      0.00164     0.000000653 0.00110       -0.569
 8 aid            0.000000748 0.000000653 0.00182       11.2
 9 air            0.00257     0.000758    0.000000657  -11.9
10 american       0.00304     0.000000653 0.00256       -0.245
# ... with 311 more rows
```

토픽 별 쌍대비교

```
# ggplot을 이용하여 beta_spread2의 그래프 생성

beta_spread2 %>%
  group_by(direction = log_ratio > 0) %>%
  top_n(10, abs(log_ratio)) %>%
  ungroup() %>%
  mutate(term = reorder(term, log_ratio)) %>%
  ggplot(aes(term, log_ratio)) +
  geom_col() +
  labs(y = "Log2 ratio of beta in topic 3 / topic 1") +
  coord_flip()
```

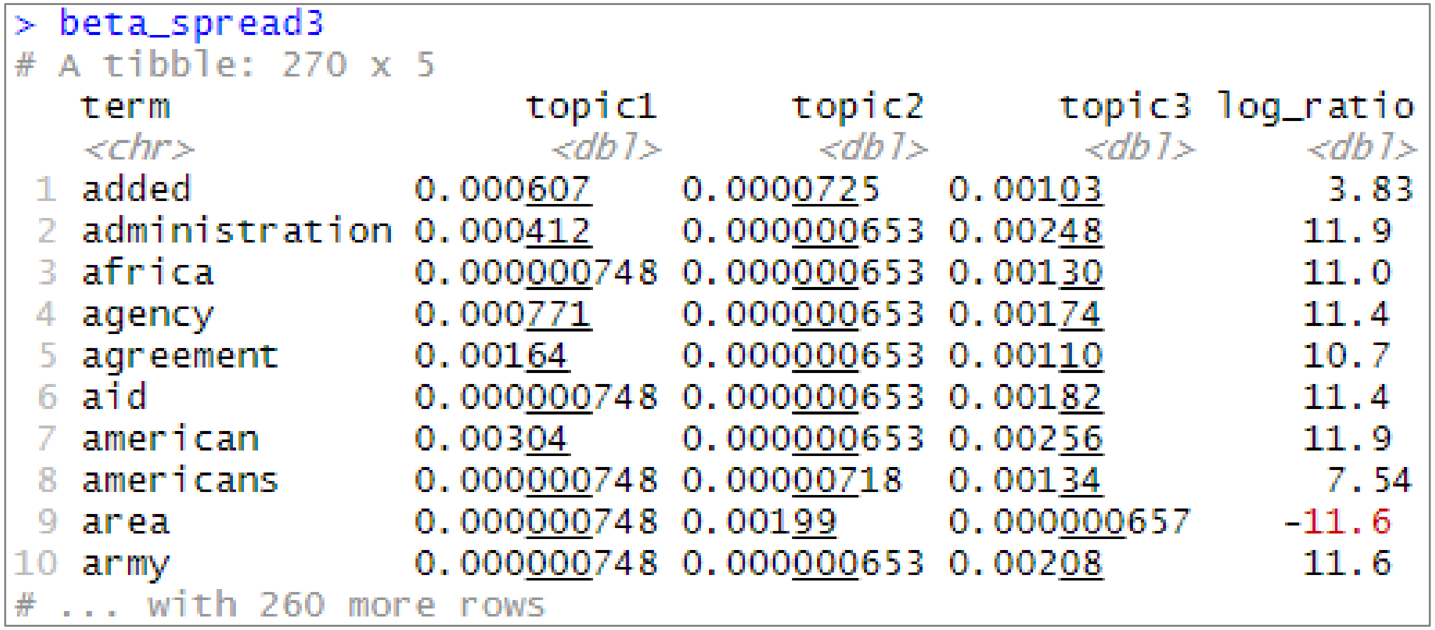
토픽 별 쌍대비교


soviet
house
military
i
south
people
union
political
told
officials
workers
stock
prices
york
federal
market
company
billion
million
percent
term
-15
-10
-5
0
5
10
Log2 ratio of beta in topic 3 / topic 1

토픽 별 쌍대비교

```
# beta_spread3 : topic 2와 topic3 비교

beta_spread3 <- ap_topics %>%
  mutate(topic = paste0("topic", topic)) %>%
  spread(topic, beta) %>%
  filter(topic2 > .001 | topic3 > .001) %>%
  mutate(log_ratio = log2(topic3 / topic2))
```



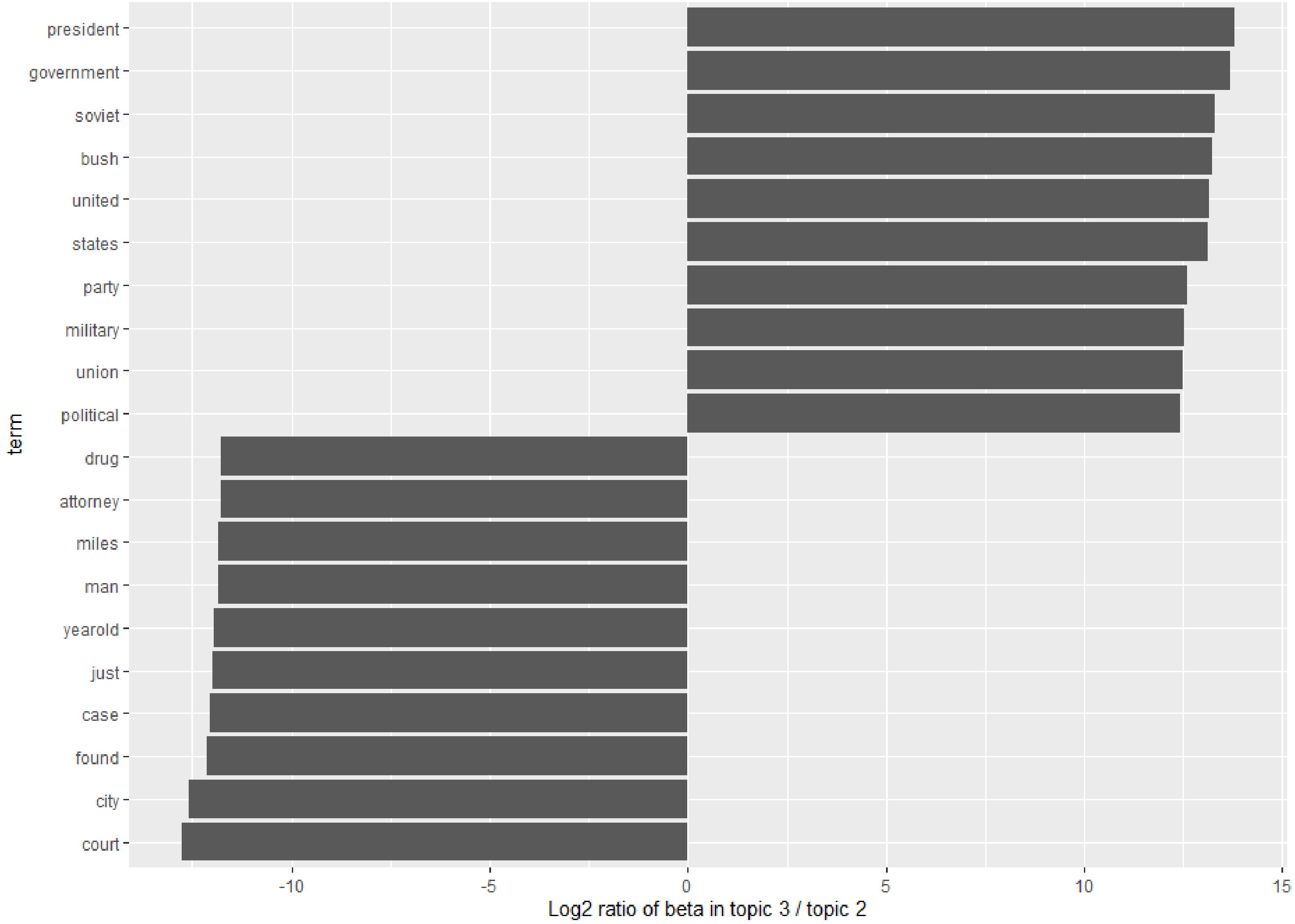
```
> beta_spread3
# A tibble: 270 x 5
   term                topic1      topic2      topic3 log_ratio
   <chr>                <dbl>       <dbl>       <dbl>     <dbl>
 1 added          0.000607    0.0000725   0.00103          3.83
 2 administration 0.000412    0.000000653 0.00248         11.9
 3 africa         0.000000748 0.000000653 0.00130         11.0
 4 agency         0.000771    0.000000653 0.00174         11.4
 5 agreement      0.00164     0.000000653 0.00110         10.7
 6 aid            0.000000748 0.000000653 0.00182         11.4
 7 american       0.00304     0.000000653 0.00256         11.9
 8 americans      0.000000748 0.00000718  0.00134          7.54
 9 area           0.000000748 0.00199     0.000000657    -11.6
10 army           0.000000748 0.000000653 0.00208         11.6
# ... with 260 more rows
```

토픽 별 쌍대비교

```
# ggplot을 이용하여 beta_spread3의 그래프 생성

beta_spread3 %>%
  group_by(direction = log_ratio > 0) %>%
  top_n(10, abs(log_ratio)) %>%
  ungroup() %>%
  mutate(term = reorder(term, log_ratio)) %>%
  ggplot(aes(term, log_ratio)) +
  geom_col() +
  labs(y = "Log2 ratio of beta in topic 3 / topic 2") +
  coord_flip()
```

토픽 별 쌍대비교


president
government
soviet
bush
united
states
party
military
union
political
drug
attorney
miles
man
yearold
just
case
found
city
court
term
-10
-5
0
5
10
15
Log2 ratio of beta in topic 3 / topic 2

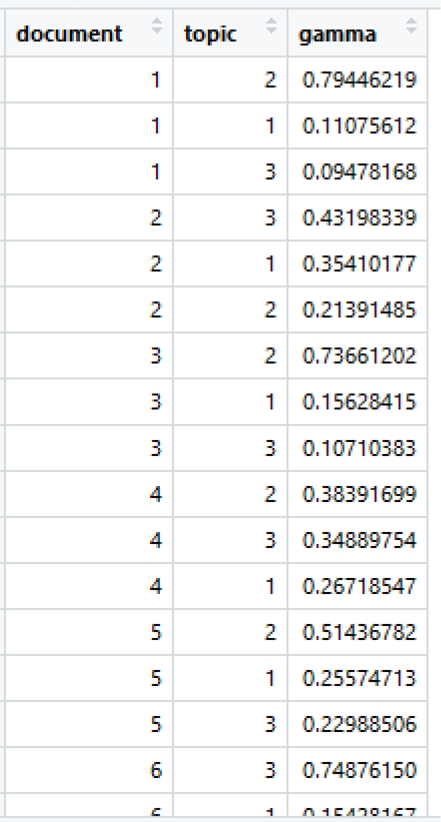
감마 탐색

```
# 구축된 모형으로부터 gamma(문서 별 토픽 확률분포) 도출

ap_documents <- tidy(ap_lda, matrix = "gamma")
ap_documents

ap_top_documents <- ap_documents %>%
  group_by(document) %>%
  top_n(3, gamma) %>%
  ungroup() %>%
  arrange(document, -gamma)

ap_top_documents
```



document	topic	gamma
1	2	0.79446219
1	1	0.11075612
1	3	0.09478168
2	3	0.43198339
2	1	0.35410177
2	2	0.21391485
3	2	0.73661202
3	1	0.15628415
3	3	0.10710383
4	2	0.38391699
4	3	0.34889754
4	1	0.26718547
5	2	0.51436782
5	1	0.25574713
5	3	0.22988506
6	3	0.74876150
6	1	0.15428167