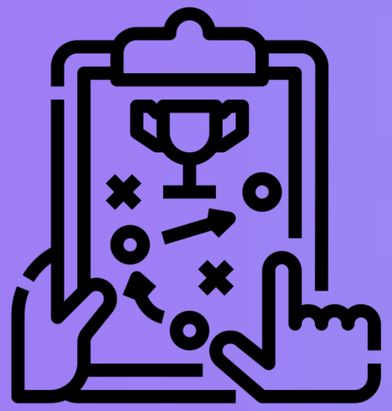
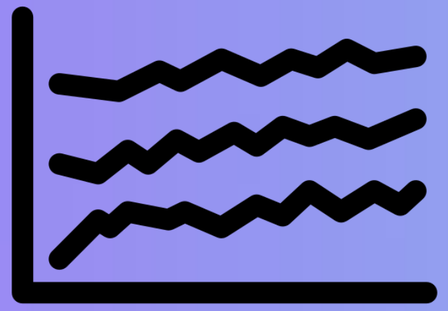
통계학과 빅데이터 페스티벌

감정과 관련된 키워드를 이용한 블로그 글 텍스트 분석

통계학과 20181504420 정해원
통계학과 20201504415 유진희





01 분석 배경 및 목표

02 데이터 전처리&EDA

03 모델링(LDA)

04 결과 해석, 결론, 한계점

01 분석 배경 및 목표

분석 배경

분석 목표

분석 배경

팀원 모두 심리와 관련된 주제에 관심이 있었고, **특히 그 중에서도 인간 내면의 우울감에 관련된 주제를 탐구하고 싶어했습니다.**

분석 목표

심리와 관련된 자료를 찾기가 어려워, 직접 저희가 텍스트 마이닝을 하게 되었습니다.

❶ 사람이 **특정 심리 상태에서 어떤 표현을 많이 쓰는지**를 분류해보는 것이 목표이며

❷ 타인의 언어 표현을 바탕으로 당시의 심리 상태를 광범위하게 나마 유추해볼 수 있도록 하고자 합니다.

02 전처리&EDA

데이터 전처리

EDA

데이터 전처리

```
for (i in 1:50){
  httr::GET(url    = "https://section.blog.naver.com/ajax
/SearchList.nhn",
            query = list("countPerPage" = "7",
                         "currentPage"  = i,
                         "endDate"      = "2022-11-23",
                         "keyword"      = "일상",
                         "orderBy"      = "sim",
                         "startDate"    = "2021-01-01",
                         "type"         = "post"),
            add_headers("referer" = "https://section.blog.naver
.com/Search/Post.nh")) %>% httr::content(as = "text") %>%
str_remove(pattern = '\\)\\]\\}\',') %>% jsonlite::fromJSON() ->
naverBlog

  data <- naverBlog$result$searchList
  Nblog.lda1 <- dplyr::bind_rows(Nblog.lda1, data)

  cat(i, "번째 페이지 정리 완료 \n")
  Sys.sleep(time=0.2)
}
```

```
#내용물 크롤링(에러 표시 코드 포함)
for(i in 1:nrow(Nblog.lda1)){
  tryCatch({
    Nblog.lda1$contents[i] <- httr::GET(url = Nblog.lda1$url[i])
%>%     xml2::read_html() %>% rvest::html_nodes(css = "div.se
-main-container")%>% html_text(trim = TRUE)
    cat(i, "번째 블로그 글 내용 취합 완료\n")
    Sys.sleep(time = 0.1)
  },
  error = function(e) cat(' --> 에러\n'))
}
```

❶ 네이버 블로그에서 **지정한 키워드 검색 후 첫 50페이지 크롤링**

❷ 한 페이지 당 게시글 7개 X 50페이지 **= 350개의 글 크롤링**

❸ **중복, 결측치, 비공개 글 확인 후 제거**

❹ 한글, 숫자, 영어, 공백 제외 **특수기호 문자 제거**

❺ 불용어 사전(list) 제작 후 **불용어 제거**

02 전처리&EDA

EDA

워드 클라우드

- 불용어 제거 (약 600개 선정) + '-다'로 끝나는 용언 제거

- 키워드: 일상

모든 단어

상위 50개 단어들

02 전처리&EDA

EDA

워드 클라우드

- 불용어 제거 (약 600개 선정) + '-다'로 끝나는 용언 제거

- 키워드: 우울증

모든 단어

상위 50개 단어들

EDA

워드 클라우드 주요 단어들

- 불용어 제거 (약 600개 선정) + '-다'로 끝나는 용언 제거

일상

사진, 운동, 집, 일기, 카페, 친구, 여행,...

우울증

생각, 사람, 우울, 병원, 약, 마음, ...

- 검색 키워드 별로 Word Cloud에서 큰 차이를 보인다
- '일상' 키워드에서는 긍정적인 단어들이, '우울증' 키워드에서는 부정적인 단어들로 이루어진 Word Could가 생성되었다
- 워드 클라우드는 단순 단어 수를 평균 내는 것에 가깝다. 더 세부적인 주제들을 찾기 위해 다른 방법을 도입해본다.

03 모델링(LDA)

1) 키워드 '우울증'
2) 키워드 '일상'
3) 키워드 '주간일기 챌린지'
4) 키워드 '행복'
5) 키워드 '사랑 '

03 모델링(LDA)

전처리 코드

```
#자연어처리2 3단계 - 체언, 용언 추출
normal.blog.ko.words <- function(text){
  pos <- str_split(text, ";") ##띄어쓰기를 기준으로 한 문장을
 여러 단어로 나눔
  pos1 <- paste(SimplePos22(pos))
  extracted <- str_match(pos1, "([가-힣a-zA-Z0-9]+)/[NC]")

  keyword <- extracted[,2]
  keyword[!is.na(keyword)]
}
```

```
#LDA - TDM
options(mc.cores=1)
normal.Corpus <- VCorpus(VectorSource(normal.blog$contents1))
normal.tdm <- TermDocumentMatrix(normal.Corpus, control=list
(tokenize=normal.blog.ko.words, removePunctuation=T,
removeNumbers=T, wordLengths=c(2,10)))
```

```
#TDM 정제
normal.tdm <- removeSparseTerms(normal.tdm, sparse=0.99)
wordFreq <- slam::row_sums(normal.tdm)
wordFreq <- sort(wordFreq, decreasing=T)
head(wordFreq)

#DTM
dtm <- as.DocumentTermMatrix(normal.tdm)
rowTotals <- apply(dtm, 1, sum)
dtm.new <- dtm[rowTotals>0, ]
dtm <- dtm.new
```

❶ 불용어 제거 후 체언, 용언만 추출

❷ TDM 생성, 정제

❸ DTM 생성하여 단어 빈도수와 결합한 행렬 생성

❹분류할 토픽 수 지정 후 LDA 시행

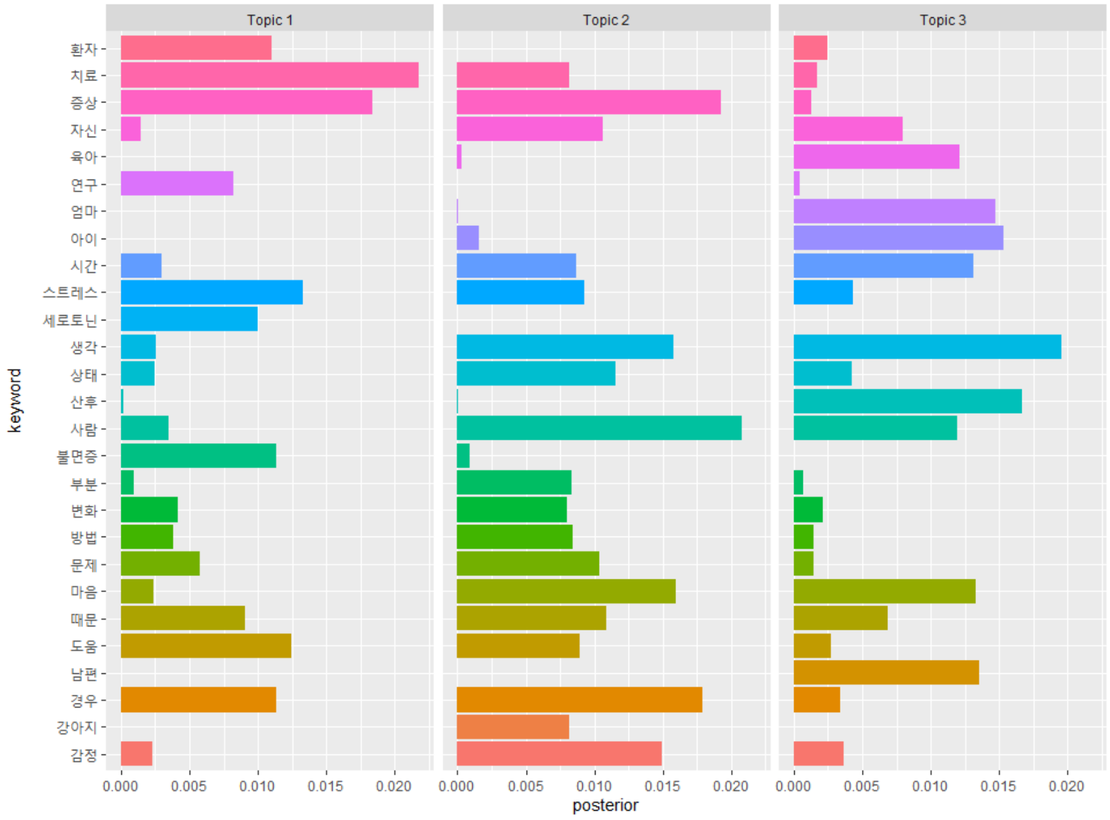
❺ 사후 확률 시각화

❻ 단어들의 사후 확률을 분석해 토픽 추측

03 모델링 (LDA)

LDA 결과들

1) 키워드 '우울증'



Topic = 3

(1) 주요 단어들

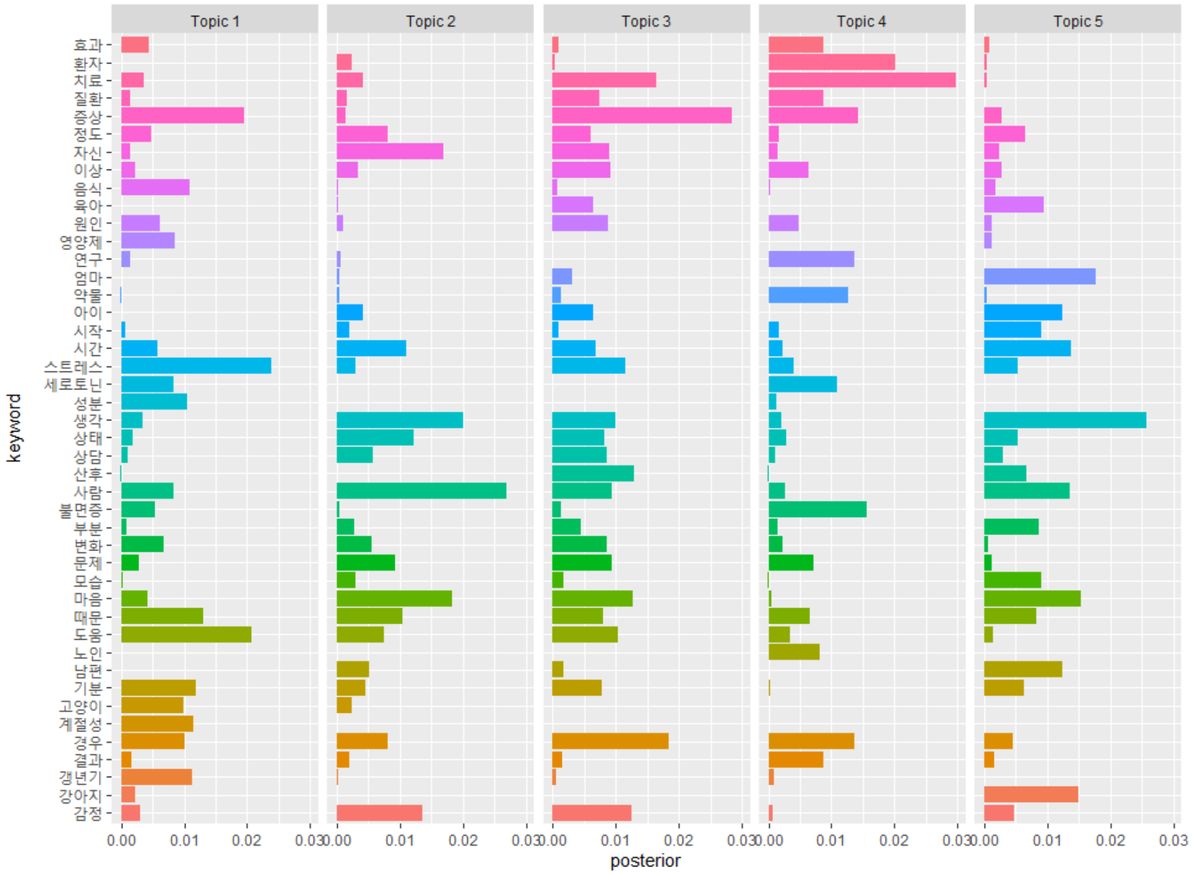
- Topic 1 : 치료, 증상, 스트레스, 도움, 경우. 환자, 세로토닌
- Topic 2 : 사람, 경우, 치료, 생각, 감정
- Topic 3 : 생각, 산후, 아이, 엄마, 마음, 도움

Topic 1	**불면증**
Topic 2	**치료후기**
Topic 3	**산후 우울증**

03 모델링 (LDA)

LDA 결과들

1) 키워드 '우울증'



Topic = 5

(1) 주요 단어들

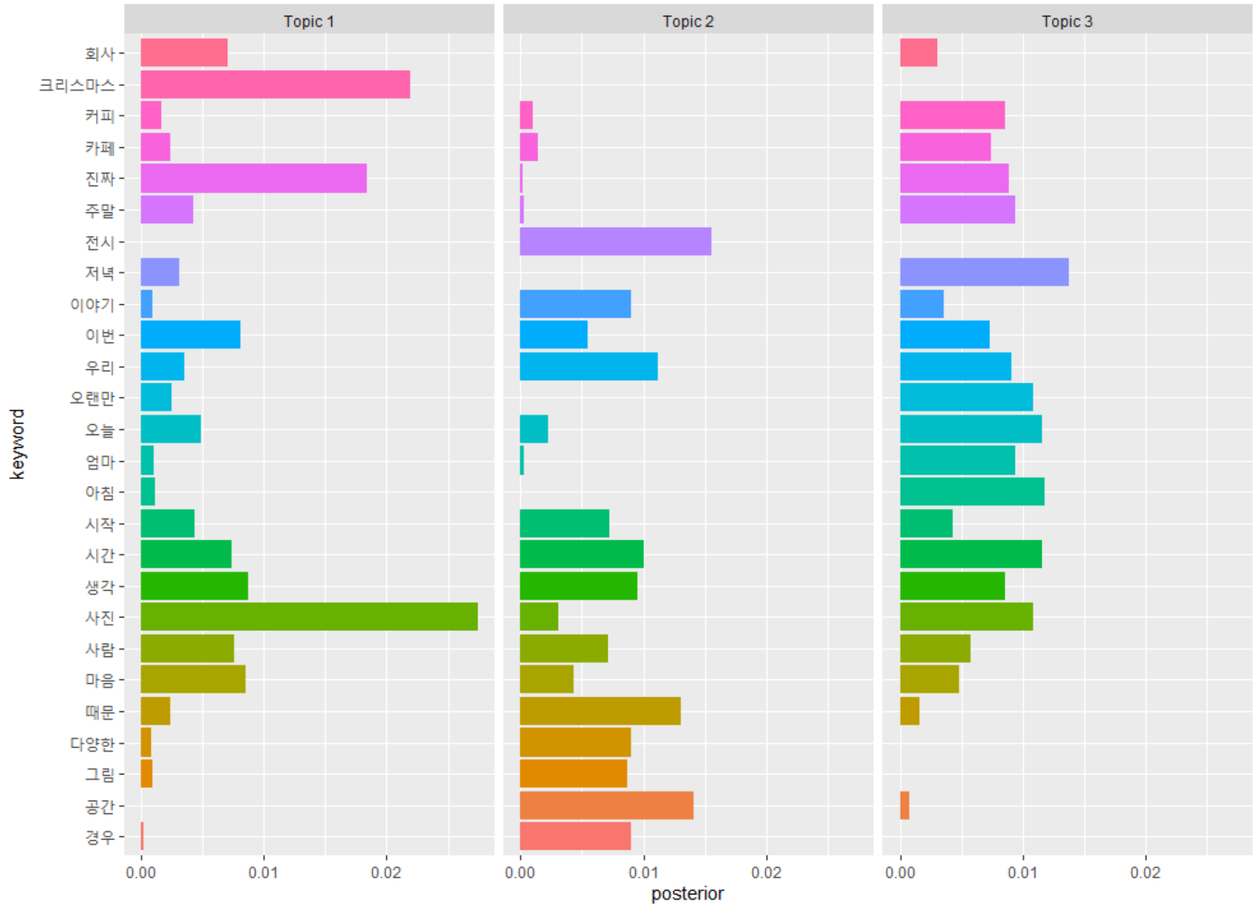
- Topic 1 : 스트레스, 도움, 증상
- Topic 2 : 사람, 생각, 마음, 자신, 상태, 감정
- Topic 3 : 증상, 치료, 경우, 마음, 산후, 감정
- Topic 4 : 치료, 환자, 불면증, 강아지, 모습
- Topic 5 : 생각, 엄마, 마음, 강아지

Topic 1	**스트레스**
Topic 2	**인간관계**
Topic 3	**산후 우울증**
Topic 4	**불면증**
Topic 5	**위로**

03 모델링(LDA)

LDA 결과들

2) 키워드 '일상'



Topic = 3

(1) 주요 단어들

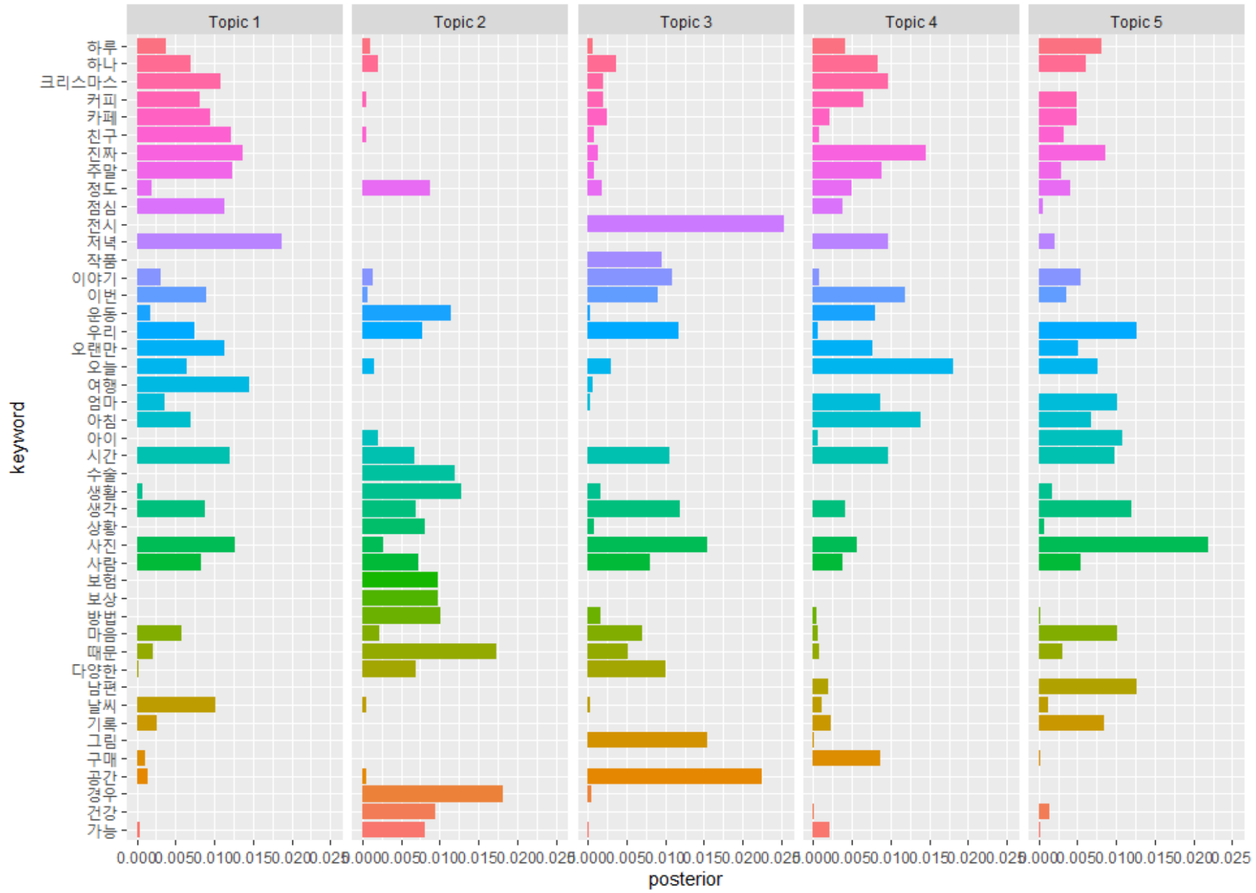
- Topic 1 : 사진, 크리스마스, 마음
- Topic 2 : 전시, 때문, 공간, 우리
- Topic 3 : 저녁, 오늘, 아침, 시간, 사진, 오랜만

Topic 1	**크리스마스, 연휴**
Topic 2	**전시, 취미**
Topic 3	**하루 일과**

03 모델링 (LDA)

LDA 결과들

2) 키워드 '일상'



Topic = 5

(1) 주요 단어들

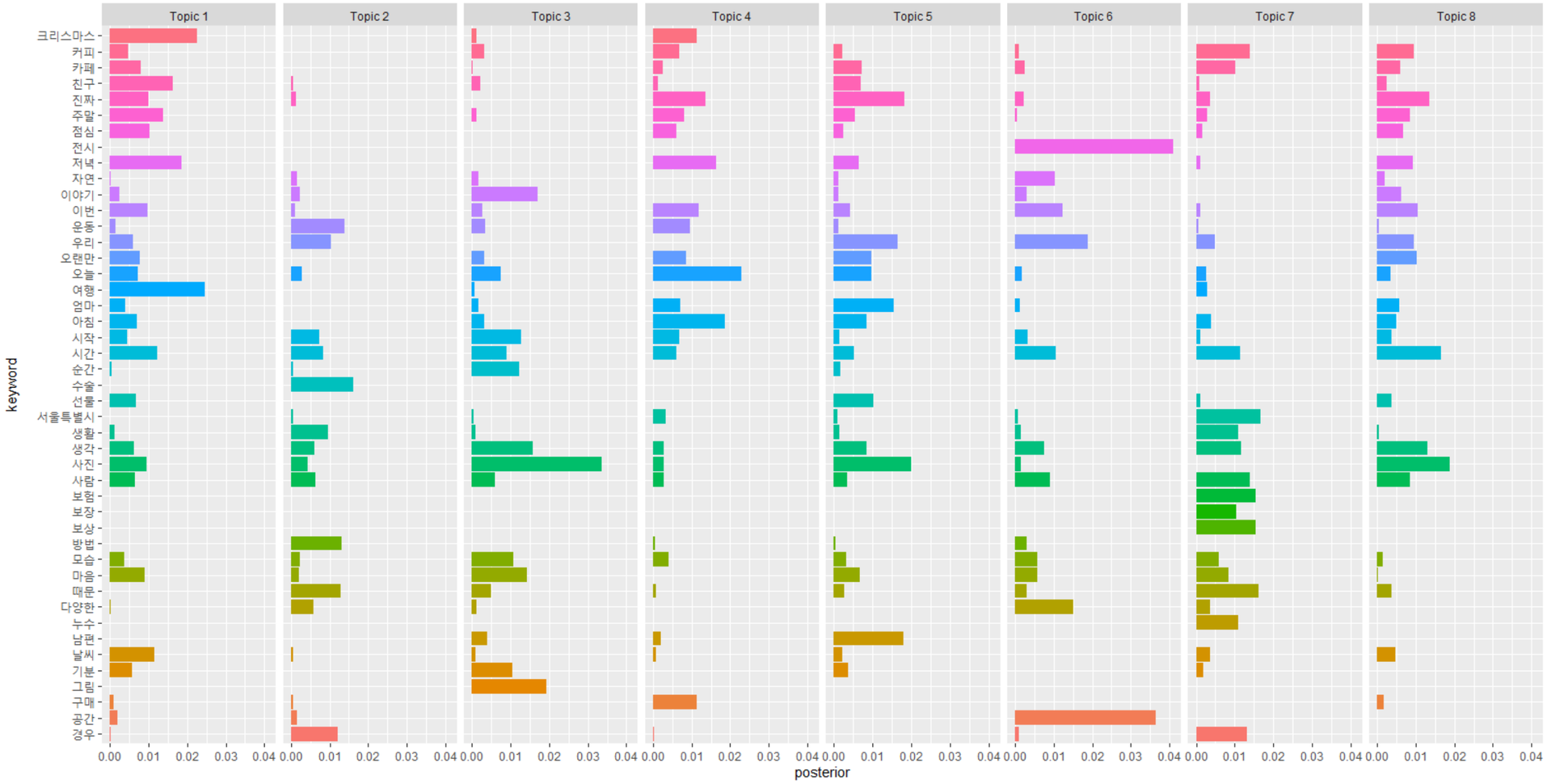
- Topic 1 : 저녁, 여행, 진짜, 주말, 친구
- Topic 2 : 운동, 생활, 수술
- Topic 3 : 전시, 공간, 그림, 사진, 생각
- Topic 4 : 오늘, 진짜, 아침, 이번
- Topic 5 : 사진, 남편, 생각, 우리

Topic 1	**우정**
Topic 2	**건강 관리**
Topic 3	**취미/ 전시**
Topic 4	**일기, 하루 일과**
Topic 5	**가족**

03 모델링(LDA)

LDA 결과들

2) 키워드 '일상'



Topic = 8

LDA 결과들

2) 키워드 '일상'

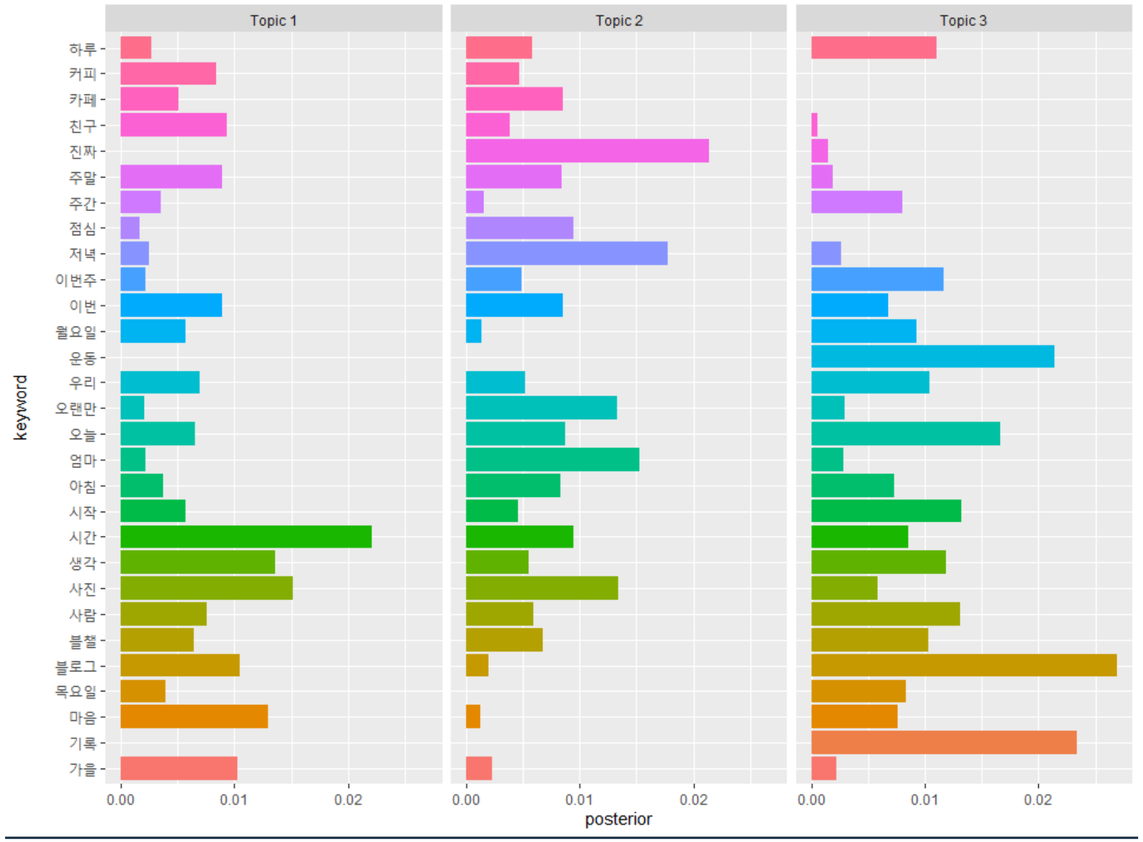
(1) 주요 단어들
- Topic 1 : 여행, 크리스마스, 저녁, 친구
- Topic 2 : 수술, 운동, 방법
- Topic 3 : 사진, 시작, 순간, 마음
- Topic 4 : 오늘, 저녁, 아침
- Topic 5 : 사진, 우리, 엄마, 남편
- Topic 6 : 전시, 공간, 우리
- Topic 7 : 커피, 서울특별시, 크리스마스
- Topic 8 : 사진, 생각, 친구

Topic 1	**연휴, 연말 기록**
Topic 2	**건강 관리**
Topic 3	**다짐, 목표**
Topic 4	**하루 일과**
Topic 5	**가족**
Topic 6	**취미, 전시회**
Topic 7	**연휴**
Topic 8	**우정**

03 모델링 (LDA)

LDA 결과들

3) 키워드 '주간일기 챌린지'



Topic = 3

(1) 주요 단어들

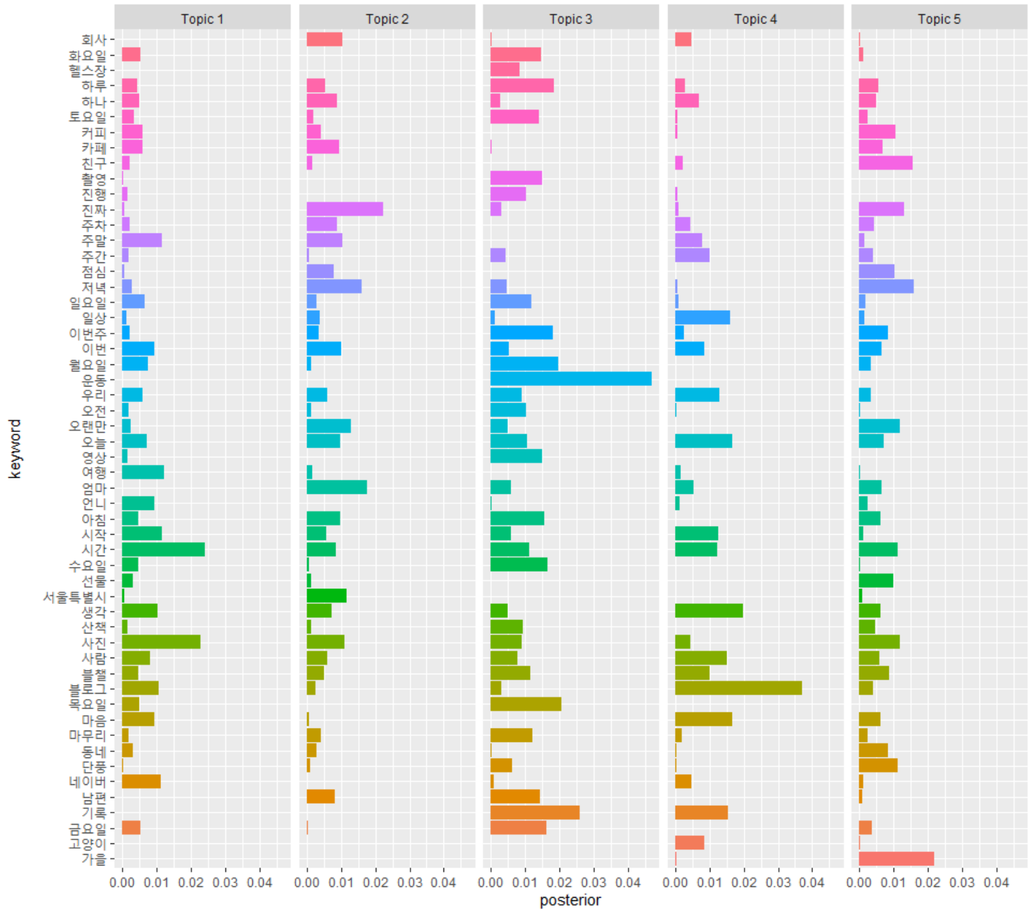
- Topic 1 : 시간, 사진, 마음, 블로그, 가을, 생각
- Topic 2 : 진짜, 저녁, 오랜만, 엄마, 사진, 시간
- Topic 3 : 블로그, 기록, 운동, 오늘, 시작, 사람

Topic 1	**기록**
Topic 2	**오랜 인연**
Topic 3	**운동, 도전**

03 모델링 (LDA)

LDA 결과들

3) 키워드 '주간일기 챌린지'



Topic = 5

(1) 주요 단어들

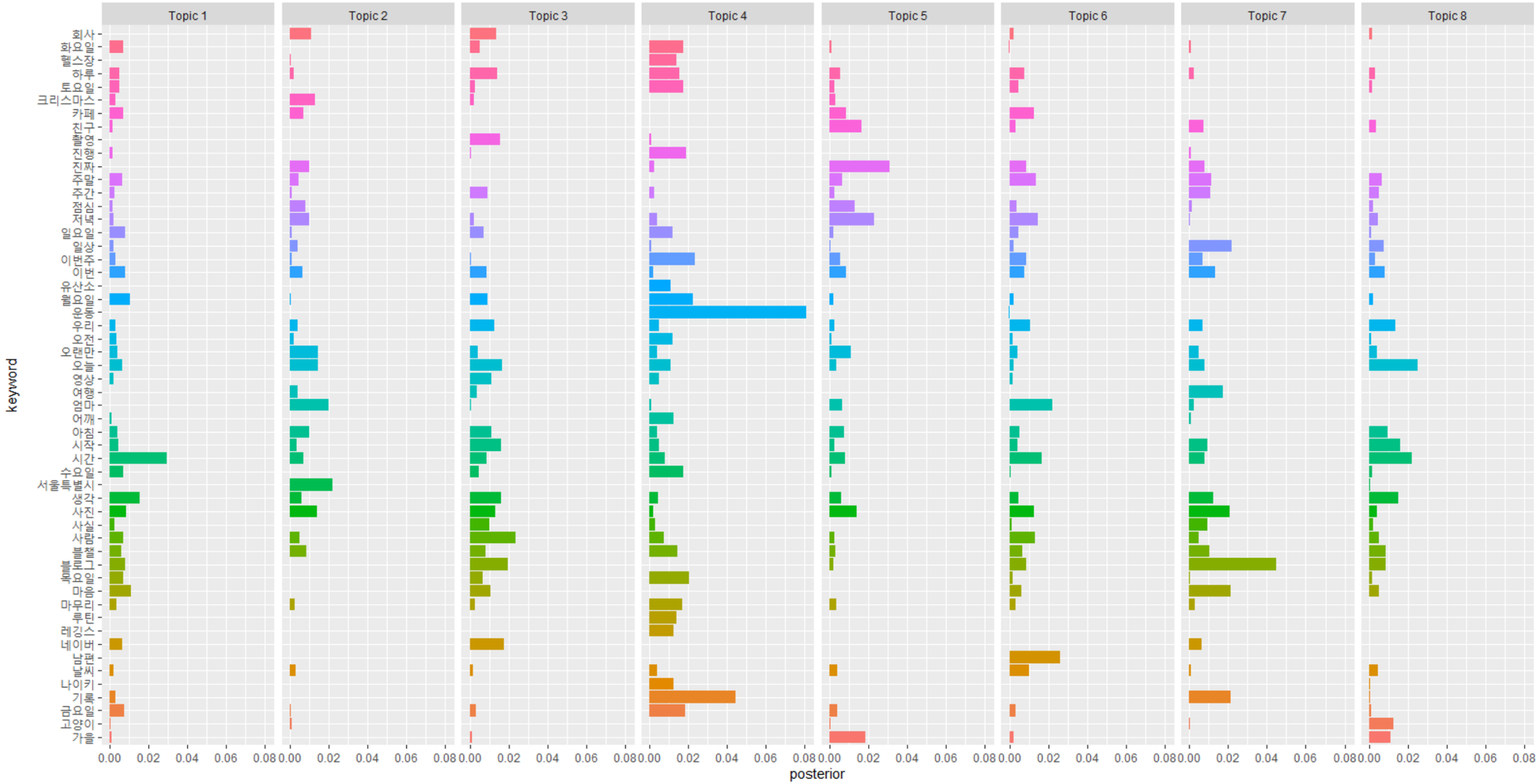
- Topic 1 : 시간, 사진, 시간, 주말, 여행, 시작
- Topic 2 : 진짜, 엄마, 서울특별시, 사진, 남편, 저녁, 오랜만
- Topic 3 : 운동, 이번주, 월요일, 기록, 금요일, 하루
- Topic 4 : 블로그, 마음, 생각, 기록, 일상, 오늘
- Topic 5 : 저녁, 오랜만, 가을, 사진, 단풍

Topic 1	**여행 기록**
Topic 2	**가족**
Topic 3	**운동**
Topic 4	**하루 일기**
Topic 5	**추억, 가을**

03 모델링(LDA)

LDA 결과들

3) 키워드 '주간일기 챌린지'



Topic = 8

LDA 결과들

3) 키워드 '주간일기 챌린지'

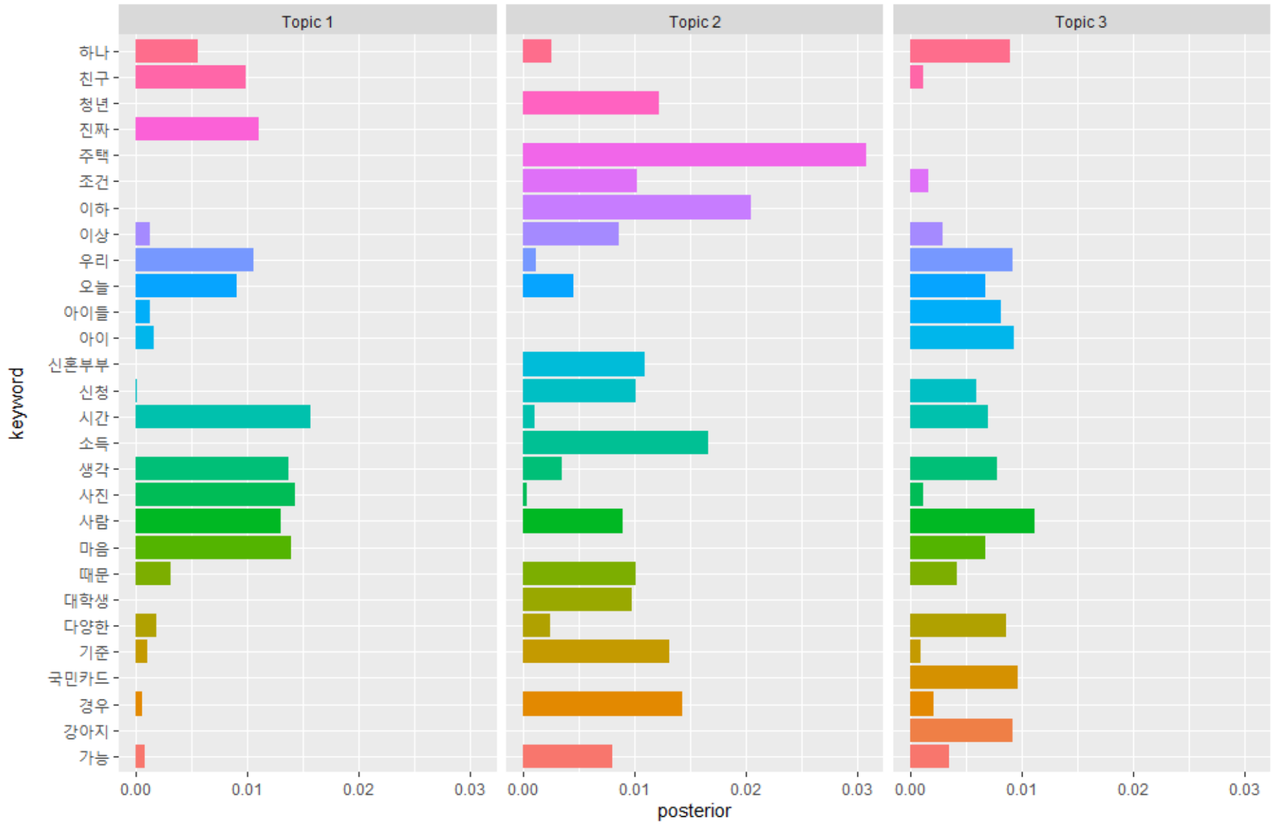
(1) 주요 단어들
- Topic 1 : 시간, 생각, 마음
- Topic 2 : 엄마, 서울특별시, 사진, 오랜만, 오늘, 크리스마스
- Topic 3 : 사람, 블로그, 네이버, 오늘, 시작
- Topic 4 : 운동, 기록, 진행, 이번주
- Topic 5 : 저녁, 가을
- Topic 6 : 엄마, 남편, 시간
- Topic 7 : 블로그, 기록, 마음
- Topic 8 : 오늘, 시간, 생각, 고양이

Topic 1	**마음 성찰**
Topic 2	**연휴**
Topic 3	**블로그 기록**
Topic 4	**운동**
Topic 5	**가을**
Topic 6	**가족**
Topic 7	**마음 성찰**
Topic 8	**마음 일기**

03 모델링 (LDA)

LDA 결과들

4) 키워드 '행복'



Topic = 3

(1) 주요 단어들

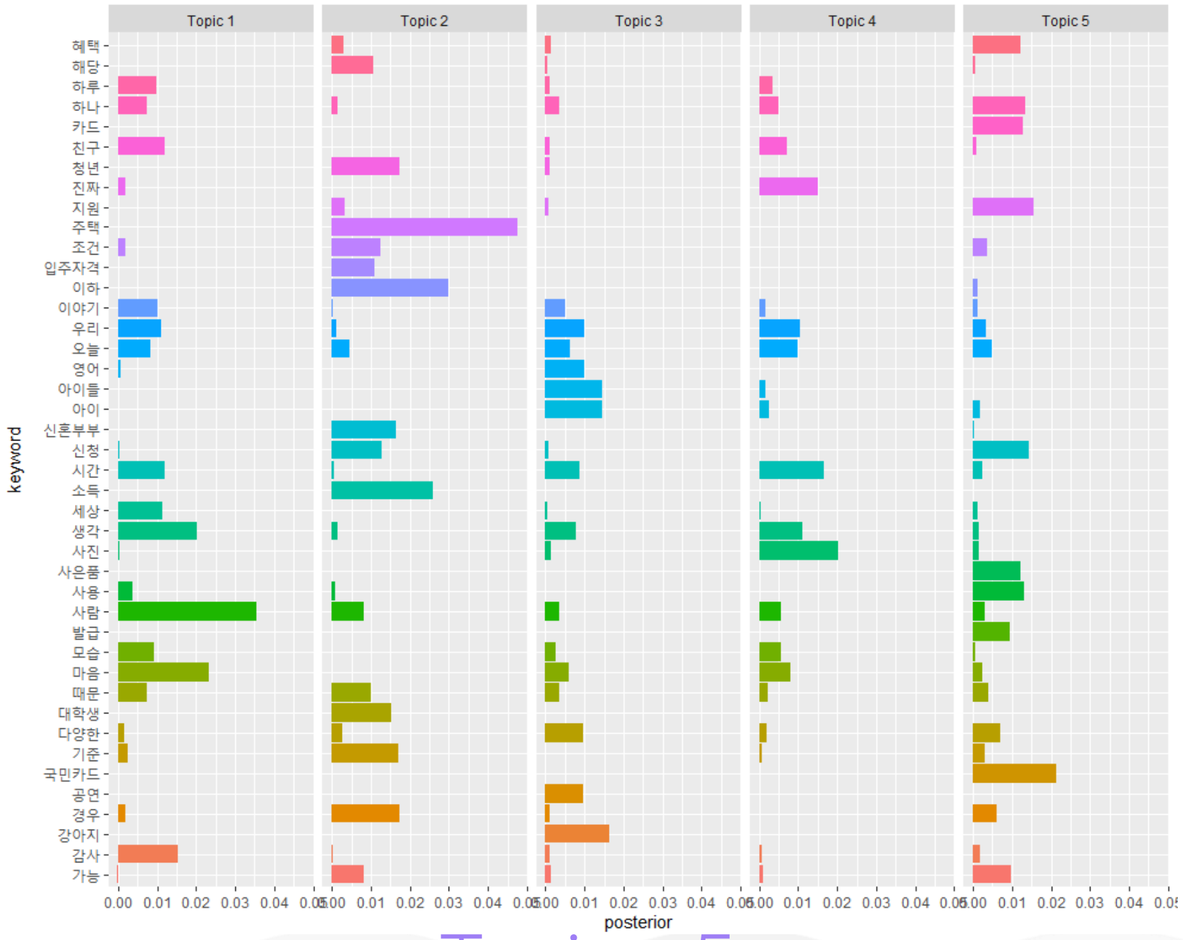
- Topic 1 : 시간, 사진, 마음, 우리, 사람, 생각
- Topic 2 : 주택, 이하, 소득, 경우, 기준
- Topic 3 : 사람, 강아지, 우리, 아이

Topic 1	**인연, 기억**
Topic 2	**물질적 행복**
Topic 3	**가족과 행복**

03 모델링(LDA)

LDA 결과들

4) 키워드 '행복'



Topic = 5

(1) 주요 단어들

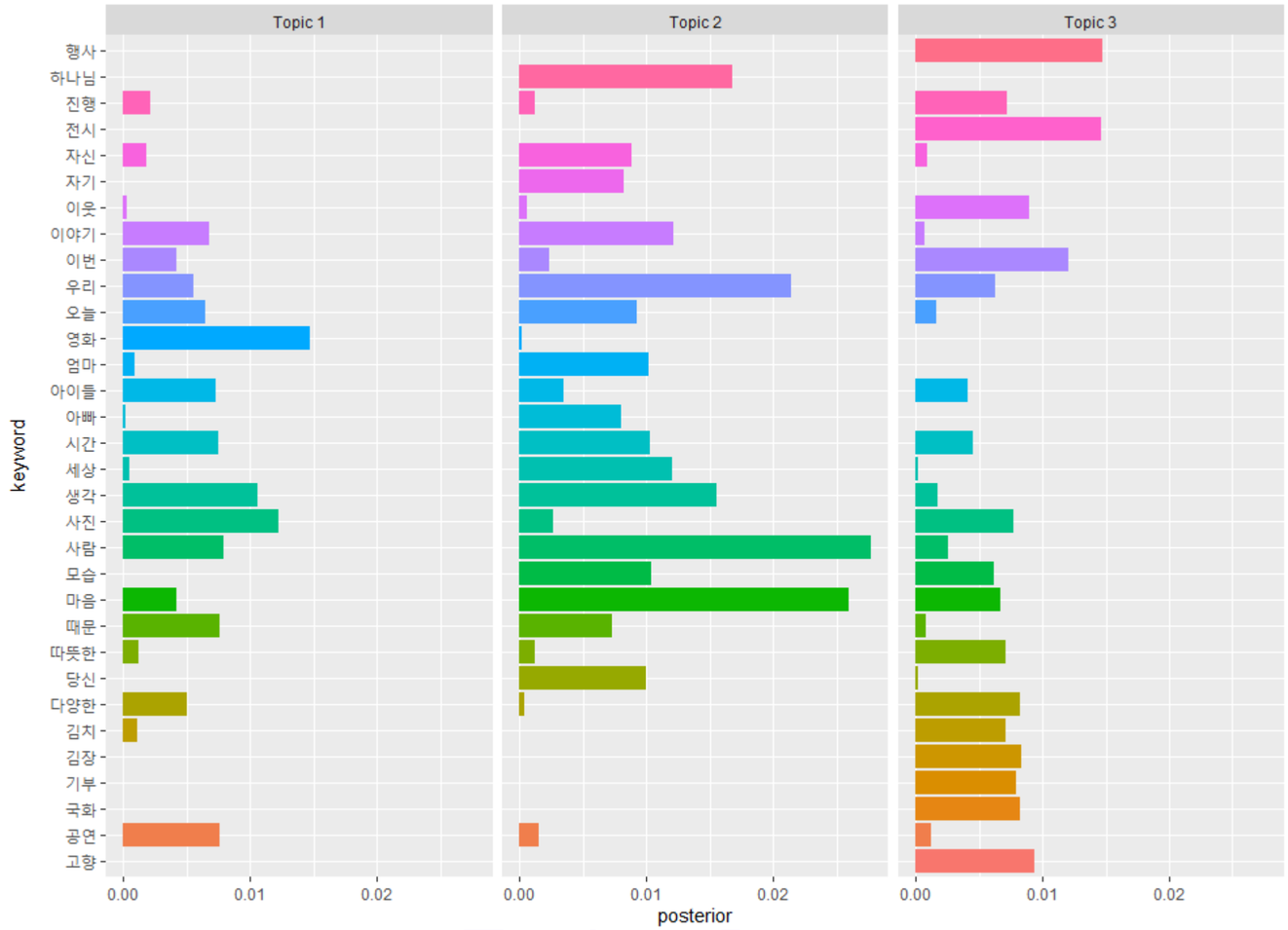
- Topic 1 : 사람, 마음, 감사, 생각
- Topic 2 : 주택, 이하, 소득, 기준, 경우
- Topic 3 : 아이들, 아이, 강아지
- Topic 4 : 사진, 시간, 우리
- Topic 5 : 지원, 기준, 신청, 사은품

Topic 1	**정신적 행복**
Topic 2	**물질적 행복**
Topic 3	**가족이 주는 행복**
Topic 4	**추억과 행복**
Topic 5	**금전적 혜택과 행복, 광고**

03 모델링(LDA)

LDA 결과들

5) 키워드 '사랑'



Topic = 3

(1) 주요 단어들

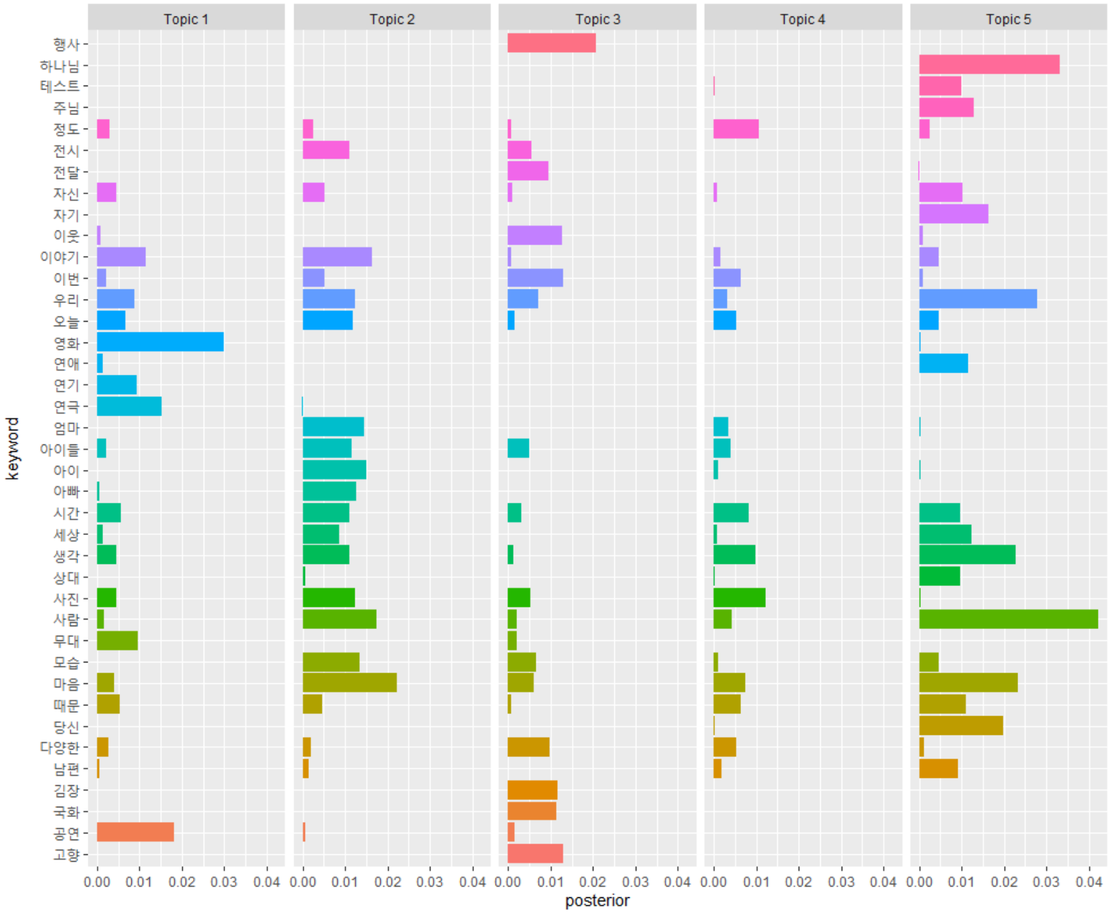
- Topic 1 : 영화, 사진, 생각, 사람, 공연
- Topic 2 : 사람, 마음, 우리, 하나님
- Topic 3 : 행사, 이야기, 전시, 고향, 기부, 김장

Topic 1	**데이트**
Topic 2	**종교적 사랑(기독교)**
Topic 3	**이웃 사랑**

03 모델링(LDA)

LDA 결과들

5) 키워드 '사랑'



Topic = 5

(1) 주요 단어들

- Topic 1 : 영화, 연극, 공연
- Topic 2 : 마음, 사람, 아이, 엄마, 이야기
- Topic 3 : 행사, 고향, 김장, 이야기,
- Topic 4 : 사진, 정도, 생각, 시간, 마음
- Topic 5 : 하나님, 사람, 마음, 당신, 생각

Topic 1	**데이트**
Topic 2	**가족**
Topic 3	**이웃사랑**
Topic 4	**사랑의 기록**
Topic 5	**종교적 사랑(기독교)**

04 결과 해석, 결론, 한계점

결과 해석

결론

한계점

결과 해석

LDA를 통한 텍스트 마이닝 및 분석을 진행하며 다음과 같은 특징들을 파악했다.

❶ Topic 수가 3일 때는 토픽 수가 더 많을 때에 비해 포괄적으로 분류가 되는 대신, 더 확실하게 분류가 된다

❷ Topic 수가 5일 때 대부분의 키워드에서 가장 깔끔하게 주제 분류가 이루어졌다

❸ Topic 수가 적정 수보다 많아진다면 오히려 토픽 간 경계가 모호해짐을 실험적으로 확인했다

결론

❶ 검색 키워드 '일상'과 '우울증'의 word cloud 결과에서 큰 차이를 보였다. 우울증이 있는 사람들은 대조군인 일상 글을 올리는 사람들에 비해 부정적인 단어를 주로 사용할 수 있음을 알 수 있었다.

❷ 가족, 취미와 같이 긍정적인 토픽들은 '행복', '일상' 등 긍정적인 키워드에 공통적으로 등장했다. 이를 통해 위 토픽들이 긍정적인 단어들과 연관되어 있음을 알 수 있다.

❸ 스트레스, 불면증 등 부정적인 느낌의 토픽들은 '우울증 ' 키워드 외에 등장하지 않았다. 즉, 위 단어들을 통해 제시한 토픽의 글을 쓰는 사람들이 우울감을 느끼고 있음을 짐작할 수 있다.

한계점

❶ R의 KoNLP 패키지 사용에 어려움이 있었습니다. KoNLP의 함수들을 정상 작동시키는데 상당히 많은 시간을 투자했습니다.

❷ 대조되는 키워드를 사용했으나, 비교적 좁은 주제를 사용하였습니다.

❸ 위 분석 결과는 어디까지나 상관 결과를 보여줄 뿐, 텍스트에 등장하는 단어로는 의학적인 진단을 할 수 없습니다.

감사합니다!