선형대수학

# 뉴스 댓글 기반 이모티콘 생성

윤지영 이지윤 이수린 임지훈 채희지

## INDEX


# 1. 주제 선정 및 배경

# 01. 주제 선정 배경

## 주제 선정


지난 4월 네이버는 뉴스 서비스에서 **감정 이모티콘을 모두 없애고** 추천 스티커로 대체

네티즌의 불만을 삼

# 01. 주제 선정 배경

## 주제 선정

요즘 네이버에서 뉴스를 끊었다. 기사를 읽다가 너무 화가 났는데 '화나요' 버튼이 없어진 것을 알고 나니 더 화가 났다.

순간적으로 '화나요' 자리의 버튼을 꾹 눌렀는데 '분석탁월' 버튼으로 바뀐 걸 알고 바로 취소했다.

정말 안타까운 일인데 이런 기사에 슬퍼요를 못 누른다는 게 어이 없네요. 쏠쏠정보, 흥미진진, 공감백배… 장난하십니까?

부고 기사에 이딴 것 밖에 누를 게 없다니, 네이버는 슬픈 유저들 감정 표현도 못 하게 하냐.

# 01. 주제 선정 배경

## 주제 선정

기사를 보고 단순히 감정 표현을 남기는 대신, 꼭 기사를 추천하고 싶을 경우 그 사유를 보여주자는 것이 취지

부정적인 의견을 낼 수 없게 해 감정 표현을 억압하는 것!
이것도 표현의 자유 침해다!
인생은 기쁨만으로는 굴러가지 않고, 슬픔이란 감정을 표현할 수 있어야 한다.
분노와 슬픔을 표출하지 못하게 막는 건 옳지 않다!

**부정적**인 의견이 다수

# 01. 주제 선정 배경

## 주제 선정

(이재진 한양대 미디어커뮤니케이션 학과 교수 의견 중)

기사에 대한 의사 표시를 하는 방식으로 감정 버튼을 애용해온 이들이

**소통 창구를 빼앗긴** 느낌이 드는 것은 당연.

네이버 측에서 오히려 **다양한 층위의 감정**을 좀 더 적극적으로 표출할 수 있도록

**대안**을 만드는 게 좋을 것!

# 01. 주제 선정 배경

## 주제 선정

네이버 기사의 특징 및 현황

### 감정분포

기사의 내용마다

전혀 다른 감정표현의 우세

### 감정표현 불균형

다양성을 반영하기에

현저히 적은 감정표현

# 01. 주제 선정 배경

## 주제 선정



네이버 기사의 특징 및 현황

감정분포

기사의 내용마다

전혀 **다른 감정표현**의 우세

감정표현 불균형

다양성을 반영하기에

누리꾼의 반응 **상이**

# 01. 주제 선정 배경

## 주제 선정

네이버 기사의 특징 및 현황


감정표현 불균형

다양성을 반영하기에

현저히 적은 감정표현

# 01. 주제 선정 배경



네이버 기사의 특징 및 현황

이미지 생성모델 활용!

**뉴스 댓글을 통해 다양한 감정을 나타내는 이모티콘 생성**

# 01. 주제 선정 배경

## 분석 흐름

**N | 주차별 분석내용**

**1주차: 각 뉴스 분야의 감정 분포 확인**

데이터 수집, 감정 분석, 욕설 탐지, LDA 진행

**3주차: 이모티콘 생성**

이미지 생성 모델

댓글 작성 → 감정분석 → 해당 감정 적합한 이모티콘 생성 →


# 2. 데이터 수집

# 02. 데이터 수집

## 기간 설정


네이버 감정표현 개편 후 사람들의 감정을
분석하는 것이 목표

이모티콘 폐지 날짜 2022년 4월 28일 이후부터 수집
댓글 수집 기간 약 6개월로 설정

*2022.05.01 ~ 2022.10.07*

## 데이터 수집 방식 설정

네이버 뉴스는 6가지 카테고리로 분류되어 있음

카테고리마다 **감정분포가 다를 것**이라 판단!

정치 ≠ 생활문화

**네이버 뉴스 6분류**에 따라 나누어 데이터 수집 진행

정치, 경제, 사회, 생활문화, 세계, 과학

# 02. 데이터 수집

## 데이터 수집 방식 설정

| 분야 | 댓글 수 | |
|---|---|---|
| 정치 | 907067 | 약 90만개 |
| 경제 | 384832 | 약 38만개 |
| 사회 | 577434 | 약 57만개 |
| 생활문화 | 177754 | 약 17 만개 |
| 과학 | 43236 | 약 5만개 |
| 세계 | 299094 | 약 29만개 |

댓글수집!

## 수집한 데이터 예시

| 분야 | 댓글 |
|---|---|
| 생활문화 | 괜히 동남아인들이 한국인을 호구취급하겠음? 우리나라는 치안 유지되는 나라라 다행이지만 명품 싸들고 비싼 고가품 들고 동남아시아 다니면 당연히 현지인들이 한국인은 돈 많다 생각해서 바가지 씌우고, 범죄피해 |
| | 저정도면 상담받아야되는거 아냐??? 오 잇.. 징그러워 |
| | 집은 없는 사람들만 상투 타령하지. 혹은 투기꾼이거나. 부동산은 주식시장이 아니야. 대폭락이고 바닥이고 없어. 부동산이 내리면 대출이자가 너무 올랐던가 경기가 개판이라 현금이 씨가 말랐다는거라서 어차피 돈없는 사람은 집 못 사. 집이 너무 비싸 보인다는 건 개나소나 대출 쉽게해준다는거거든. 실수요자면 일단 하나 사야 나중에라도 오를때 남들 자산늘어나는거 보면서 울지 않지. |
| | 돈이 남으면 그돈으로 국민한테 주던지 쓸데없는생각하고 낭비하니까 지지율떨어지지 탄핵이야기도 나오는판국이구만 한심하다 그리고 대통령이란 사람이 굽신거리고 인사하냐? 격떨어진다 |
| | 지금 먹다가, 냄새가 이상해서 찾아보니- 이런 기사가ㅜㅜㅋㅋ |


# 3. 감정분류

# 03. 감정분류

## 분류된 감정 확인

AI HUB의 '한국어 감정 정보가 포함된 단발성 대화 데이터셋'

| | A | B |
|---|---|---|
| 32828 | 서동욱 격하게 응원한다 화이팅 아낌없이 실력발휘해요 | 행복 |
| 32829 | 머리 ㅋㅋㅋㅋㅋㅋㅋ캐릭터 있음ㅋㅋㅋㅋㅋ | 행복 |
| 32830 | 서래마을 부부 너무 인품도 좋으시고 교양도 넘치시고 보는내내 부럽고 행복해보여서 좋았습니다 ^^ | 행복 |
| 32831 | 유해진씨 삼시세끼에봤는데 간만에 복귀작이네요 코미디라서 재밌겠다 꼭보러가야지ㅋㅋ | 행복 |
| 32832 | 위대한 국민, 위대한 대한민국 | 행복 |
| 32833 | 21세기판 팔만대장경ㅋㅋㅋㅋㅋㅋㅋㅋㅋ 지린다 | 행복 |
| 32834 | 응원합니다...칡즙도 간절히기다리구여..사랑합니다♥♥♥ | 행복 |
| 32835 | 간만에 레전드급 무도 봤다 ㅋㅋㅋ | 행복 |
| 32836 | 진해수 우투ㄱ ㅋㅋㅋㅋ | 행복 |
| 32837 | 박한이 선수 내년에도 100안타 기다립니다 | 행복 |
| 32838 | 은근 꿀잼이던데~~~ | 행복 |
| 32839 | 후반기부터 마무리 맡았는데도 기세가 대단하네. | 행복 |
| 32840 | 히트다히트 | 행복 |
| 32841 | 대한민국은 부끄럽지만 대한민국 국민이라는것은 자랑스럽습니다 | 행복 |
| 32842 | 박형식 삼맥종 캐릭터 최고! | 행복 |

SNS 글 및 온라인 댓글에 대한 웹 크롤링을 통해 선정한 약 4만 개의 문장

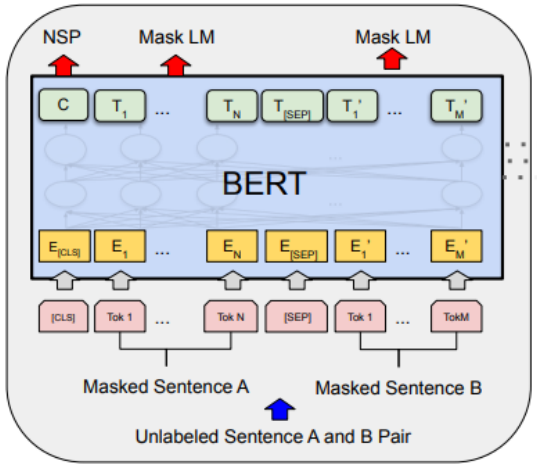
**7개의 감정(기쁨, 슬픔, 놀람, 분노, 공포, 혐오, 중립)으로 같은 수**만큼 라벨링

## BERT란?

BERT

Transformer의 인코더 부분 사용

대량의 단어 임베딩 등 사전학습이 되어있는 모델




Pre-training




Fine Tuning

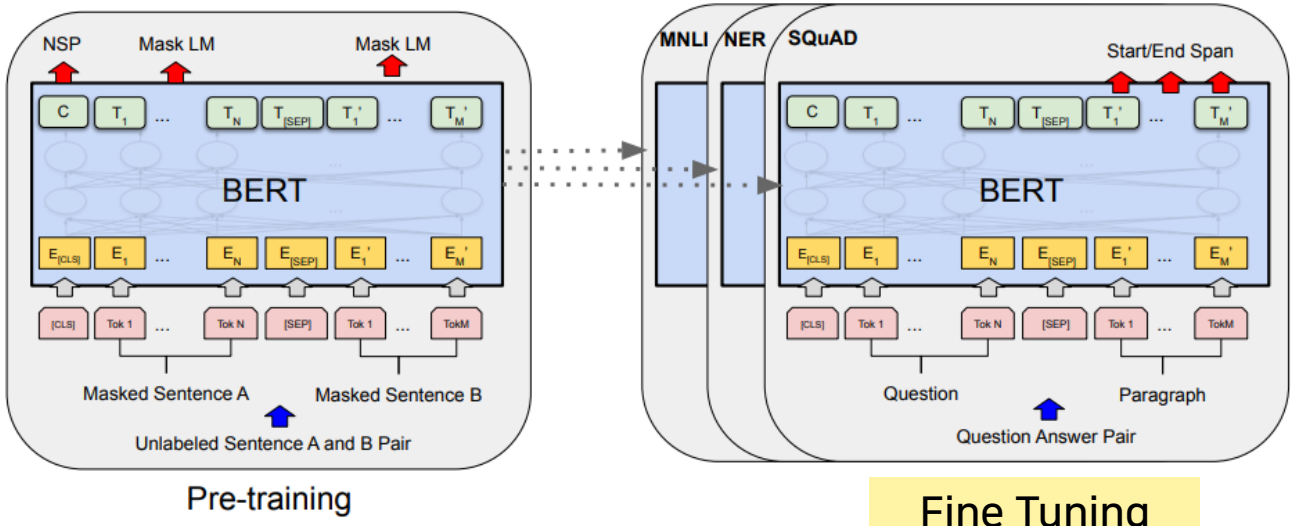
추가적으로 원하는 task에 맞춰 추가학습 가능!

## BERT란?

BERT

Transformer의 인코더 부분 사용

대량의 단어 임베딩 등 사전학습이 되어있는 모델



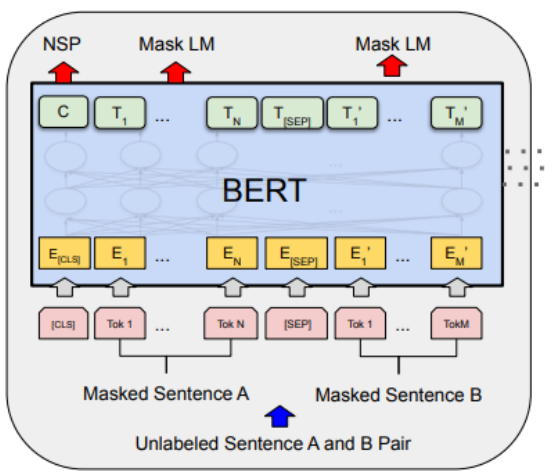

원하는 task에 맞춰 추가학습 가능!

## BERT란?

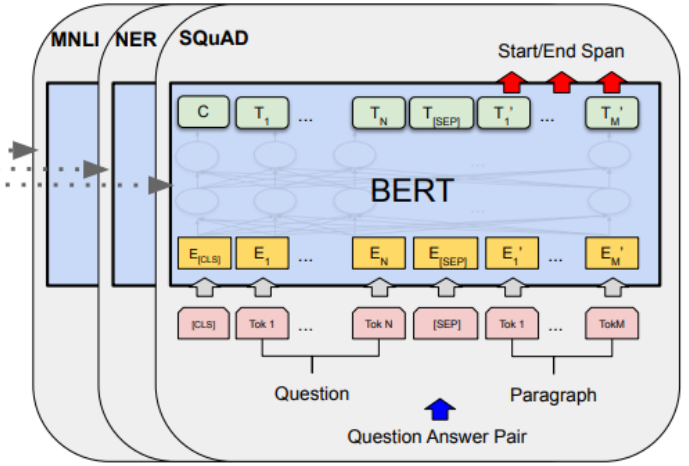
BERT

Transformer의 인코더 부분 사용

대량의 단어 임베딩 등 사전학습이 되어있는 모델

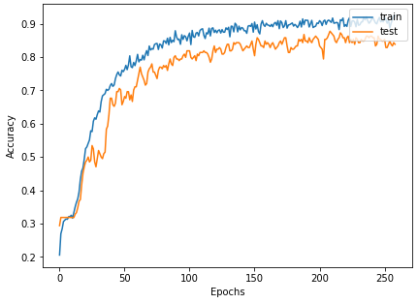



Pre-training




Fine Tuning

Test acc: 0.81273…

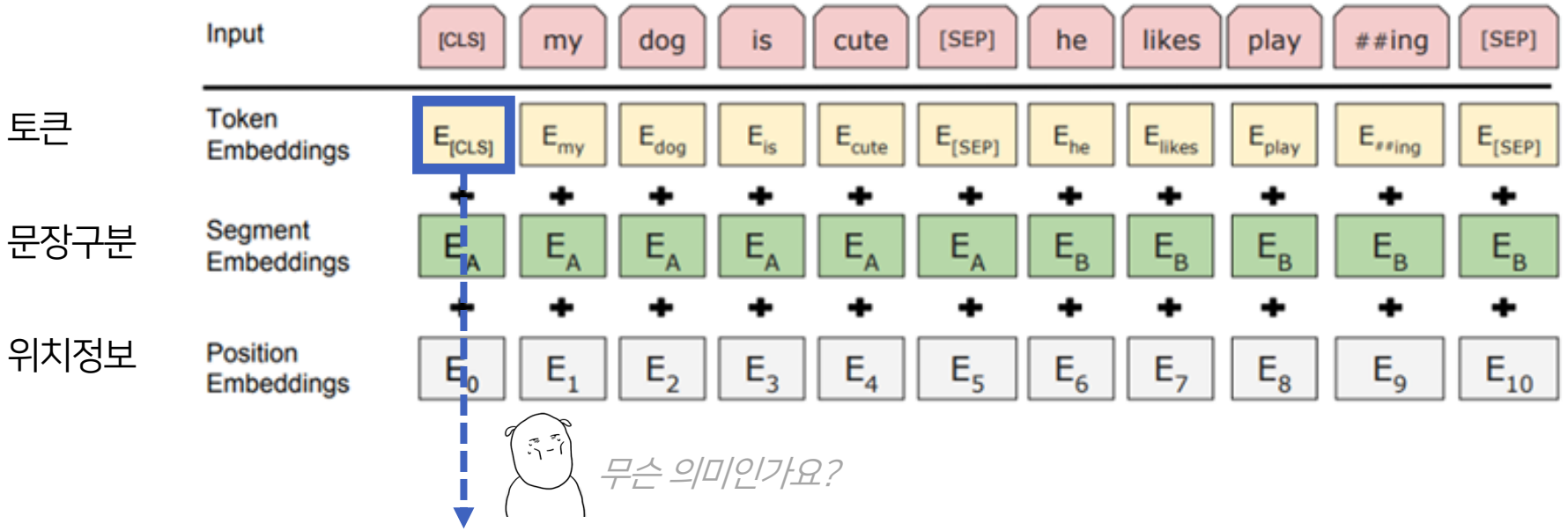



AIhub데이터로

Fine Tuning 진행

원하는 task에 맞춰 추가학습 가능!

## BERT의 input에 필요한 임베딩 벡터




CLS 토큰: 입력 받은 모든 문장의 시작

전체 계층을 다 거친 후 토큰 시퀀스의 결합된 의미 ▪▪▪▪▪▪▪▪▪▪ 분류문제에 사용!

# 03. 감정분류

## KoBERT란?

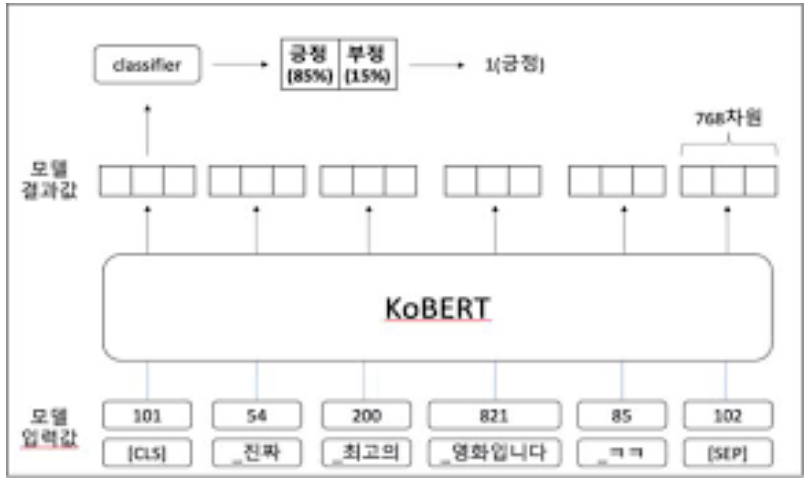
이 모델 사용!

KoBERT

SKBrain에서 개발

BERT 모델에서 한국어 데이터를 추가적으로 학습

한국어 위키에서 5백만개의 문장과 54만개의 단어를 학습시킨 모델



# 03. 감정분류

## 댓글 시각화

감정 대분류에 따른 7가지 감정

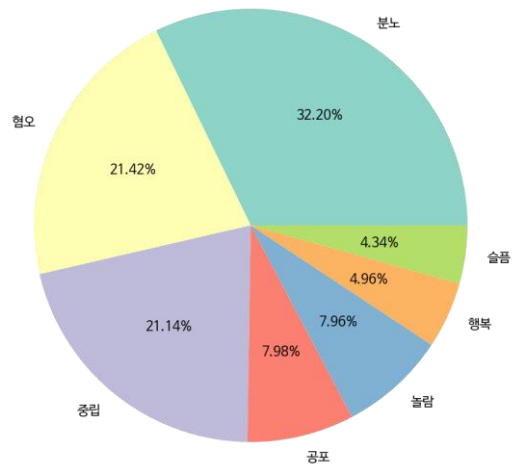
기쁨, 슬픔, 놀람, 분노,
혐오, 공포, 중립

각 분야의 댓글 별 감정 상태를 **대분류**에 따라 나누고
시각화 진행

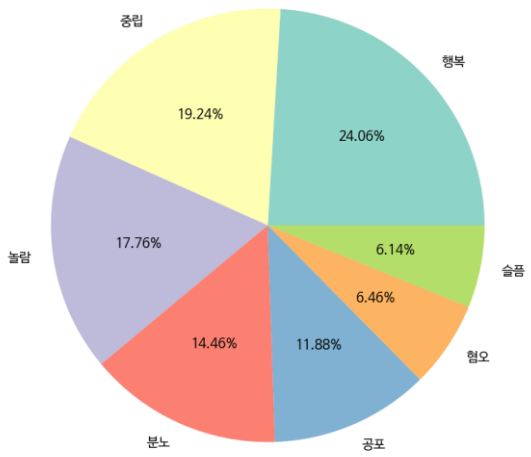
# 03. 감정분류

## 분야별 감정분류 시각화

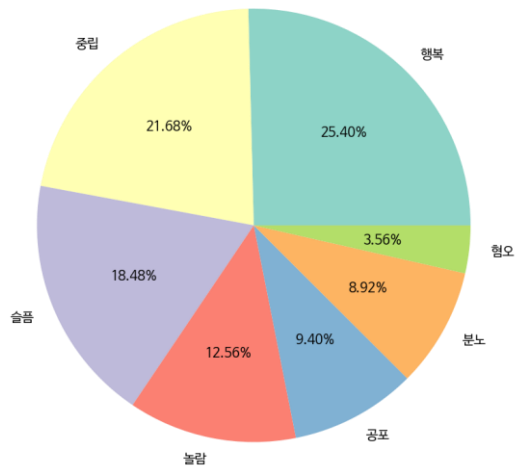
### 경제




### 과학




### 생활문화

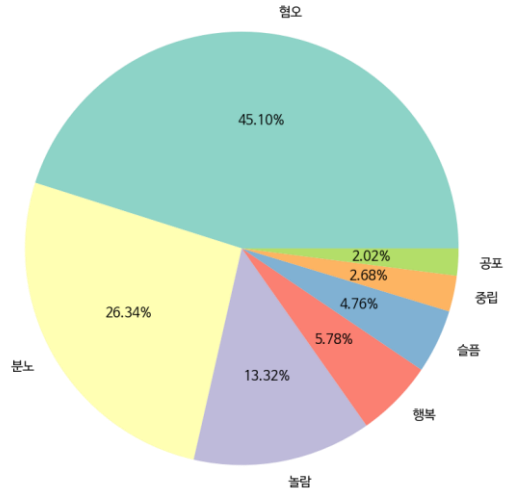



과학과 생활문화에서는 부정적인 감정과 긍정적인 감정이 **비교적 고르게** 분포 되어 있음

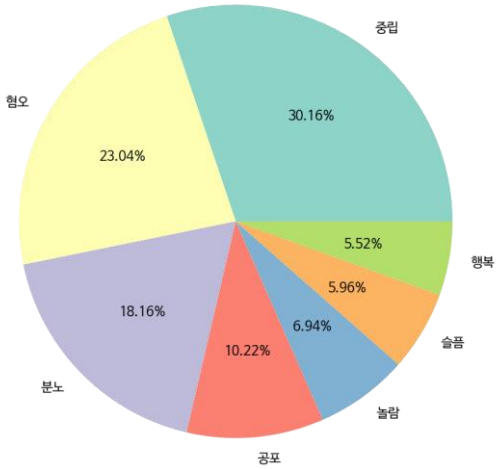
# 03. 감정분류

## 분야별 감정분류 시각화

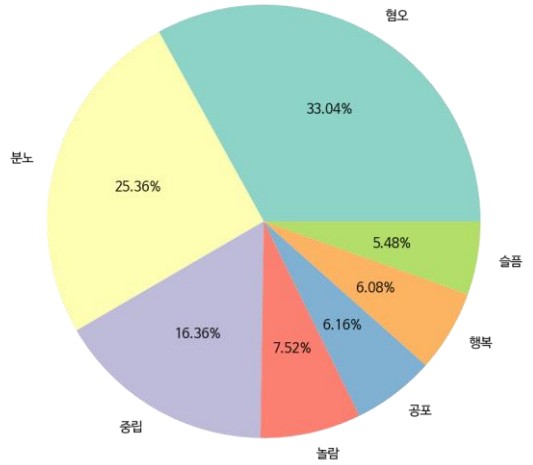
정치




세계




사회




혐오와 분노가 모든 감정의 절반 이상을 차지함

**감정의 불균형**이 심함


# 4. 욕설탐지

# 04. 욕설탐지

## 분류된 감정 확인

*결과를 한 번 확인…*

| | comment | emotion |
|---|---|---|
| 1 | 뉴진스의 인기를 보며 나이가 들었다는게 확 느껴진다 | 중립 |
| 2 | 코로나 풀려서 친구들이나 가족들과 오붓하게  식사하고 싶어도 요즘 물가에는 선뜻  약속 잡기가 힘든것 같습니다ㅜ | 슬픔 |
| 3 | 시간버려가면서 3천원벌겠다고 ㅋㅋㅋㅋ 진짜 특급상거지네 ㅋㅋ | 놀람 |
| 4 | 포레스트검프 부터 보고 봐야지 | 중립 |
| 5 | 분명 한국을 알리고 좋은일은 햇으네 오히려 공인으로써 모범이 되야되지않을까?스티브 유인지 머시깽이 유인지도 요 | 중립 |
| 6 | 강원도 너무너무 좋지 아 속초 가고 싶다. | 행복 |
| 7 | 글쓰는 좋은 팁 알려 주셔서 감사합다. | 행복 |
| 8 | 전생에 나라 10번은 구하셨나보다 축하합니다 | 행복 |
| 9 | 이눔의.교회.다.없애라 | 분노 |
| 10 | 정부는 해결은 하고 있는건지 | 분노 |
| 11 | 두방이냐 세방이냐가 관건 | 중립 |
| 12 | 광현아 그러는거 아니야 | 놀람 |
| 13 | 괜히 동남아인들이 한국인을 호구취급하겠음? 우리나라는 치안 유지되는 나라라 다행이지만 명품 싸들고 비싼 고가품 | 혐오 |
| 14 | 저정도면 상담받아야되는거 아냐???오 잇..  징그러워 | 혐오 |
| 15 | 방탄소년단의 앞으로의 활동도 기대되고 항상 응원합니다 이번 앨범 너무 좋아요~ | 행복 |
| 16 | 석열이가 숟가락들고 기다리겠네 | 중립 |
| 17 | 손꼽아 기다리던 솔로앨범이 왔네요 | 행복 |
| 18 | 인천 부개역 부개고가 점검요망 | 중립 |
| 19 | 아무리 자유라도 지켜야할 선이 있는 것이다. 그럼 대형마트 골목상권 침해한다고 불평해서도 안되는 것이다. | 분노 |
| 20 | 활동잠정전 상장시켜 돈 끌어모으고 이제서 나몰라라 | 분노 |
| 21 | 쿨링포크가 깨끗한 물을 특수 노즐을 통해 미세한 크기의 인공안개로 분사하는 설비군요. 불쾌감도 없고 주변 온도를 | 중립 |

**욕설**이 담긴 댓글의 감정을 **정확히 분류하지 못하는 문제점** 발견!

# 04. 욕설탐지

## 문제점 발견

**욕설**이 담긴 댓글의 감정을 **정확히 분류하지 못하는 문제점** 발견!

### 데이터셋 자체의 문제

혐오를 드러내는 내용임에도
'ㅋㅋㅋ'가 많아
'행복'으로 **잘못 라벨링된** 경우 多

데이터셋에 없는
**새로운 혐오표현**을 인식하지 못함

### 혐오와 중의적 표현이 함께 사용된 경우

좌빨놈 → 혐오
좌빨놈들 덕에 나라 잘 돌아간다 → 중립

미친것 → 혐오
미친것이 날뛰니 Y 지지율 올라가겠다^^
→ 행복

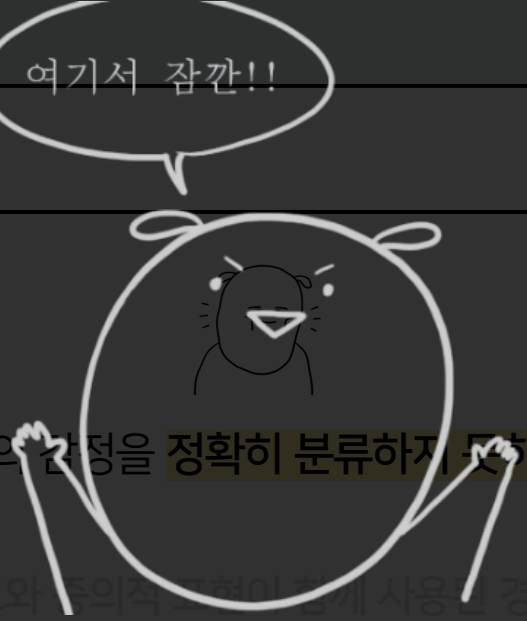
### 욕설의 형태를 변형하여 쓴 경우

뻘개이들 특징 손절 잘하는거 Y도
그렇고 K도 그렇고 조만간 빵 ㄱ ㄱ ㄱ
→ 중립

입 뻥긋만해도 정치권에 조우우웃 되는
정치인들ㅉ
→중립

## 04. 욕설탐지

### 문제점 발견




**욕설**이 담긴 댓글의 감정을 **정확히 분류하지 못하는 문제점** 발견!

데이터셋 자체의 문제

혐오를 드러내는 내용임에도

'ㅋㅋㅋ'가 많아

'행복'으로 **잘못 라벨링된** 경우 多

데이터셋에 없는

**새로운 혐오표현**을 인식하지 못함

**내용상 욕설이 다수 등장할 수 있으니 양해 바랍니다**

**선형대수학팀의 의견과 무관하며 개인의 자유로운 의사표현을 존중합니다**

# 04. 욕설탐지

## 문제점 발견

**욕설**이 담긴 댓글의 감정을 **정확히 분류하지 못하는 문제점** 발견!

### 데이터셋 자체의 문제

혐오를 드러내는 내용임에도
'ㅋㅋㅋ'가 많아
'행복'으로 **잘못 라벨링된** 경우 多

데이터셋에 없는
**새로운 혐오표현**을 인식하지 못함

### 혐오와 중의적 표현이 함께 사용된 경우

좌빨놈 → 혐오
좌빨놈들 덕에 나라 잘 돌아간다 → 중립

미친것 → 혐오
미친것이 날뛰니 Y 지지율 올라가겠다^^
→ 행복

### 욕설의 형태를 변형하여 쓴 경우

뻘개이들 특징 손절 잘하는거 Y도
그렇고 K도 그렇고 조만간 빵 ㄱ ㄱ ㄱ
→ 중립

입 뻥긋만해도 정치권에 조우우웃 되는
정치인들ㅉ
→중립

# 04. 욕설탐지

## 문제점 발견

**욕설**이 담긴 댓글의 감정을 **정확히 분류하지 못하는 문제점** 발견!

### 데이터셋 자체의 문제

혐오를 드러내는 내용임에도
'ㅋㅋㅋ'가 많아
'행복'으로 **잘못 라벨링된** 경우 多

데이터셋에 없는
**새로운 혐오표현**을 인식하지 못함

### 혐오와 중의적 표현이 함께 사용된 경우

좌빨놈 → 혐오
좌빨놈들 덕에 나라 잘 돌아간다 → 중립

미친것 → 혐오
미친것이 날뛰니 Y 지지율 올라가겠다^^
→ 행복

### 욕설의 형태를 변형하여 쓴 경우

뻘개이들 특징 손절 잘하는거 Y도
그렇고 K도 그렇고 조만간 빵 ㄱ ㄱ ㄱ
→ 중립

입 뻥긋만해도 정치권에 조우우웃 되는
정치인들ㅉ
→중립

# 04. 욕설탐지

## 문제점 발견

특정 대상을 비난하는 의도의 비속어 표현을 사용했다는 점에서

혐오(부정적) 감정에 더욱 가까울 것이라 판단

추가적으로 욕설탐지를 진행하여

**오분류된 부정적 댓글**을 찾기로!

# 04. 욕설탐지

## 전처리

**인터넷 욕설표현의 특징 고려 필요!**

필터링을 피하기 위해 욕설의 형태를 바꾸는 경우

(중성 바꾸기, 받침 바꾸기 등)

konlpy + word2vec 조합 사용 불가

konlpy와 word2vec의 자세한 내용은

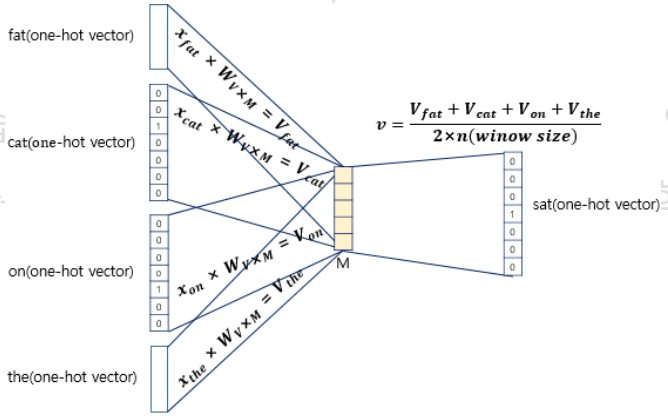
딥러닝팀3주차 클린업 참고!

# 04. 욕설탐지

## KoNLPy와 Word2Vec란?

KoNLPy : 한국어 형태소 라이브러리

Word2Vec: 단어를 벡터로 변환하는 방법 (e.g. skipgram)






# 04. 욕설탐지

## 전처리

**인터넷 욕설표현의 특징 고려 필요!**

필터링을 피하기 위해 욕설의 형태를 바꾸는 경우

(중성 바꾸기, 받침 바꾸기 등)

konlpy + word2vec 조합 사용 불가

*word2vec*

1. 학습되지 않은 단어(OOV)에 대해서 벡터 값 얻을 수 X
2. 잘 사용되지 않는 단어(Infrequent words) 학습 어려움

konlpy와 word2vec의 자세한 내용은

딥러닝팀3주차 클린업 참고!

# 04. 욕설탐지

## 전처리



인터넷 욕설표현의 특징 고려 필요!

필터링을 피하기 위해 욕의 형태를 바꾸는 경우

(중성 바꾸기, 초성 바꾸기 등)



단어를 글자 수준(Character-Level)으로

쪼개서 학습하는 임베딩 모델 FastText 사용

1. 학습되지 않은 단어(OOV)에 대해서 벡터 값 얻을 수 X
2. 잘 사용되지 않는 단어(Infrequent words) 학습 어려움

konlpy와 word2vec의 자세한 내용은

딥러닝팀3주차 클린업 참고!

## FastText

*N-gram: n개의 연속적인 단어 나열*

Fast Text

단어를 글자 수준(Character-Level)으로 쪼개서 학습하는 임베딩 모델

단어를 N-Gram의 Bag-of-Words로 표현

| 단어 | N | BoW |
|---|---|---|
| 선형대수학 | 2 | {<선, 선형, 형대, 대수, 수학, 학>, <선형대수학>} |
| | 3 | {<선형, 선형대, 형대수, 대수학, 수학>, <선형대수학>} |

N의 범위를 설정하여 사용

보통 3~6

전체 단어도 학습!

## FastText

*N-gram: n개의 연속적인 단어 나열*

Fast Text

단어를 글자 수준(Character-Level)으로 쪼개서 학습하는 임베딩 모델

단어를 N-Gram의 Bag-of-Words로 표현

| 단어 | N | BoW |
|---|---|---|
| 선형대수학 | 2 | {<선, 선형, 형대, 대수, 수학, 학>, <선형대수학>} |
| | 3 | {<선형, 선형대, 형대수, 대수학, 수학>, <선형대수학>} |

*벡터 값 다 합산하여 사용*

$$U_{\text{선형대수학}} = z_{<\text{선}} + z_{\text{선형}} \cdots + z_{\text{수학}} + z_{\text{학}>}$$

$$+ z_{<\text{선형}} + z_{\text{선형대}} \cdots + z_{\text{대수학}} + z_{\text{수학}>} + z_{<\text{선형대수학}>}$$

# 04. 욕설탐지

## 데이터 설명

| | |
|---|---|
| 좌배 까는건 ㅇㅂ | 1 |
| 집에 롱 패딩만 세 개다. 10년 더 입어야지 ㅋㅋ | 0 |
| 애새끼가 초딩도 아니고 ㅋㅋㅋㅋ | 1 |
| 재앙이한건했노 | 1 |
| 글쓴이 와꾸 승리에 비하면 방사능 파폭 원숭이 일듯… | 1 |
| 세탁이라고 바도 된다 | 0 |
| 은행에 대출 상담 받으러 가보면 직업의 귀천 바로 알려줌 | 0 |

Data

뉴스 댓글, 각종 커뮤니티 사이트의 댓글에 대해
**욕설 여부를 0과 1**로 분류한 데이터

*분류 기준*

욕설, 인종 차별, 정치적 갈등 조장, 성적·성차별적, 타인을 비하,
그 외에 불쾌감을 주거나 욕설로 판단되는 말

## 데이터 설명

'존맛', '개이득' 등의 말처럼 악의가 없는 **단순 강조의 의미**로 판단되는 경우,

**상황에 따라 의미가 달라지는 단어**는 비욕설로 구분

Data

뉴스 댓글, 각종 커뮤니티 사이트의 댓글에 대해

**욕설 여부를 0과 1**로 분류한 데이터

*분류 기준*

욕설, 인종 차별, 정치적 갈등 조장, 성적·성차별적, 타인을 비하,

그 외에 불쾌감을 주거나 욕설로 판단되는 말

## 금칙어 사전

같은 맥락!

금칙어 사전 방법을 사용하지 않은 이유

목적: 단순 욕설 필터링이 아니라 **부정적 감정이 담긴 혐오표현**을 찾자

① 동음이의어가 욕설로 분류되는 상황 방지 (Ex. 시발점)

② 욕설에는 해당하지만 부정적 의미를 담지 않은 단어 배제

*(데이터 라벨링 과정에서 이를 고려한 데이터 사용)*

③ 수많은 변형 형태 존재

## 금칙어 사전

같은 맥락!

금칙어 사전 방법을 사용하지 않은 이유

목적: 단순 욕설 필터링이 아니라 **부정적 감정이 담긴 혐오표현**을 찾자

① 동음이의어가 욕설로 분류되는 상황 방지 (Ex. 시발점)

② 욕설에는 해당하지만 부정적 의미를 담지 않은 단어 배제

*(데이터 라벨링 과정에서 이를 고려한 데이터 사용)*

③ 수많은 변형 형태 존재

→ 금칙어 사전보다는 문맥에 따라 혐오표현을 탐지해야!

# 04. 욕설탐지

## 전처리

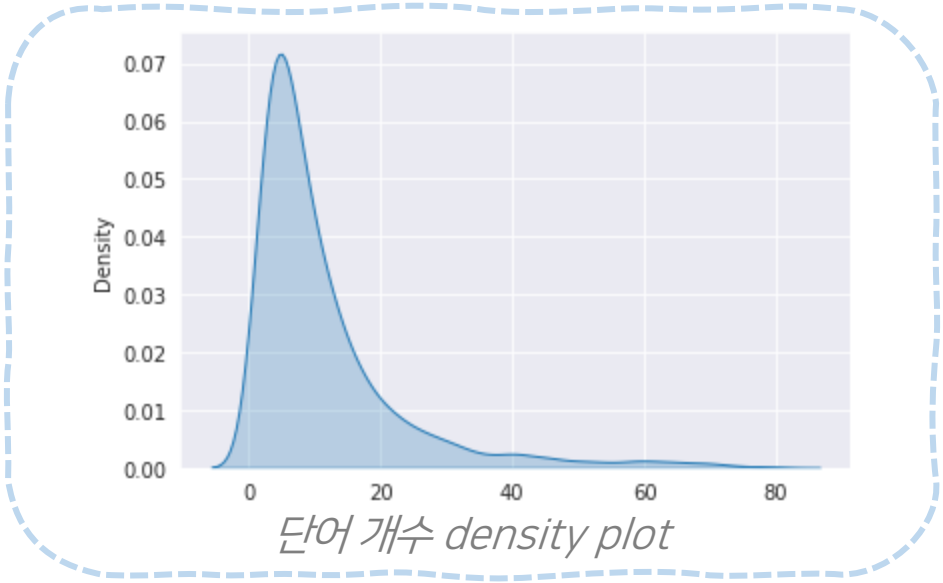
1. 특수문자, 이모티콘,반복되는 단어 제거

| 댓글 |
| --- |
| 네이버 멋집니다👏👏👏👏 |
| 일본 중국을 가지를 않는데 몬 🐕 소랴 |
| 마룬5 한국공연 취소해라~~~!!! |
| 개상청 날씨 또 바꾸네ㅋㅋㅋ |

| 댓글 |
| --- |
| 네이버 멋집니다 |
| 일본 중국을 가지를 않는데 몬 소랴 |
| 마룬5 한국공연 취소해라 |
| 개상청 날씨 또 바꾸네 |

## 전처리

2. 한 문장 내 2단어 미만/25단어 초과 문장 제거




*단어 개수 density plot*

긴 문장을 **띄어쓰기 없이** 이어서 쓴 경우 학습에 방해

→ 2단어 미만 문장 제거

최대 단어 개수 25로 통일

## 전처리

### 3. 초 · 중 · 종성 분리

변형된 단어나, 오탈자가 많은 글의 형태적 정보를 효과적으로 학습하기 위해

**자모음 수준**으로 분리

(종성 없는 경우 '_'으로 대체)

EXAMPLE

가격보다 맛있어서 가는곳 →

ㄱㅏ_ㄱㅕㄱㅂㅗ_ㄷㅏ_ ㅁㅏㅅㅇㅣㅆ
ㅇㅓ_ㅅㅓ_ ㄱㅏ_ㄴㅡㄴㄱㅗㅅ

## 전처리


3. 초 [illegible]성 분리

**왜 분리하나요…?**

변형된 단어나, 오탈자가 많은 글의 형태적 성보를 효과적으로 학습하기 위해

**형태가 조금 달라져도 유사한 n-gram을 갖기 때문!**

(종성 없는 경우 '_'으로 대체)

Ex) 선형대수학 & 선형대슈학

**많은 개수의 동일한 n-gram 학습**

## 전처리

### 4. FastText 모델을 활용한 임베딩

학습 방식 : Skip-Gram *(중심 단어로부터 주변 단어를 예측)*

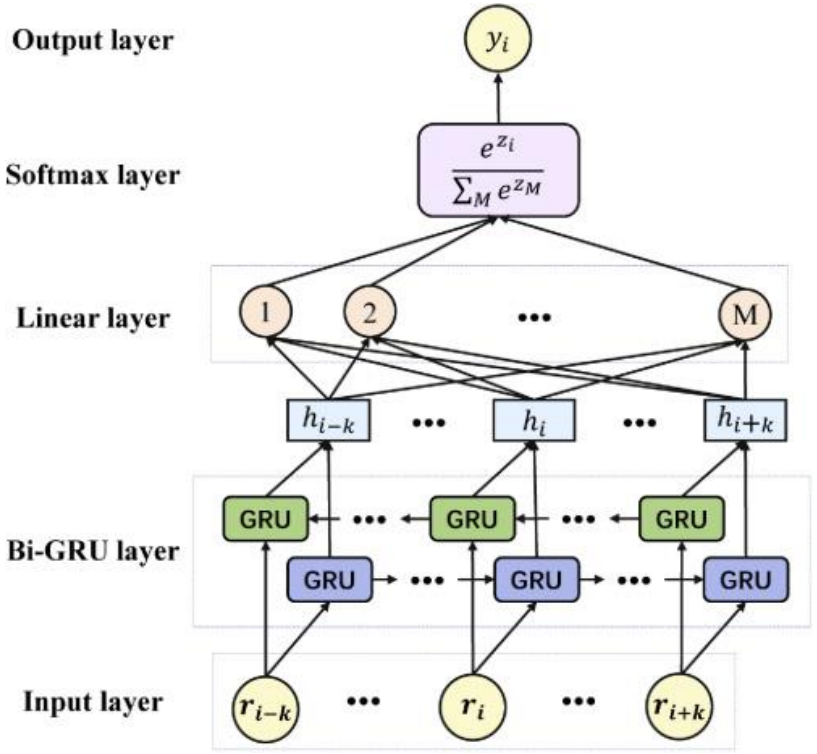
N-Gram : 3~6

| 단어 | N | BoW |
|---|---|---|
| 선형대수학 | 3 | <ㅅㅓ, ㅅㅓㄴ, ㅓㄴㅎ, …, _ㅎㅏ,ㅎㅏㄱ, ㅏㄱ> |
| | 4 | <ㅅㅓㄴ, ㅅㅓㄴㅎ, …, _ㅎㅏㄱ,ㅎㅏㄱ> |
| ⋮ | | |
| 선형대수학 | 6 | <ㅅㅓㄴㅎㅕ, …, ㅜ_ㅎㅏㄱ> |

*Skip-Gram에 대한 자세한 내용은*
*딥러닝팀3주차 클린업 참고!*

# 04. 욕설탐지

## 모델학습




Bidirectional GRU

온라인 텍스트의 **특정 단어를 탐지**하는 task의 경우,

문자 단위 임베딩+bi-lstm/gru

구성이 고성능을 낸 사례들이 많다는 연구결과 참고

*GRU의 자세한 내용은*

*딥러닝팀3주차 클린업 참고!*

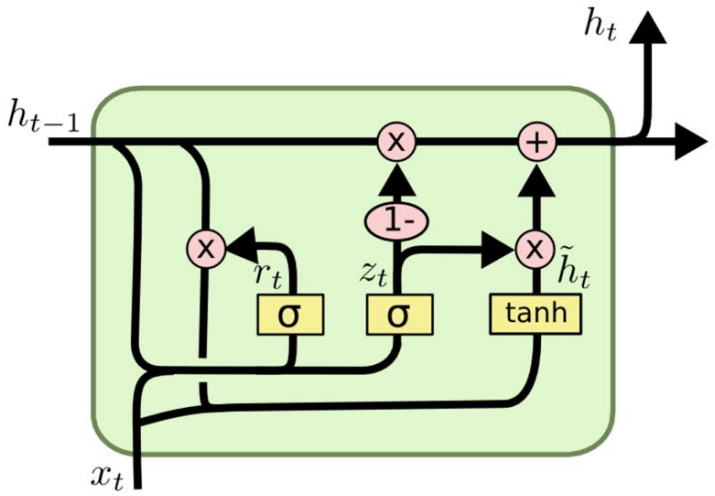
## GRU

### GRU

LSTM의 장기 의존성 문제 해결책 유지

은닉 상태 업데이트 연산량 ↓

업데이트 게이트+리셋 게이트




*GRU의 자세한 내용은*

*딥러닝팀3주차 클린업 참고!*

## Bidirectional GRU

기존 LSTM/GRU는 t시점을 분석하기 위해

t-n시점의 데이터를 사용하는 **단방향** 학습

이전에 나타난 데이터뿐만 아니라,

**이후에 나타난 데이터가** 영향을 줄 수도 있지 않을까?

## Bidirectional GRU

EXAMPLE



기존 LSTM/GRU 는 t시점을 분석하기 위해

t-n시점의 데이터를 사용하는 **단방향** 학습

나는 ___을 뒤집어 쓰고 펑펑 울었다.

이 시점에서 "나는 ___" 라는 데이터로 **빈칸 추론 불가능**

이전에서 나타난 데이터뿐만 아니라,

**이후에 나타난 데이터가** 영향을 줄 수도 있지 않을까?

## Bidirectional GRU

EXAMPLE

나는 ___을 뒤집어 쓰고 펑펑 울었다.

이 시점에서 "나는 ___" 라는 데이터로 **빈칸 추론 불가능**

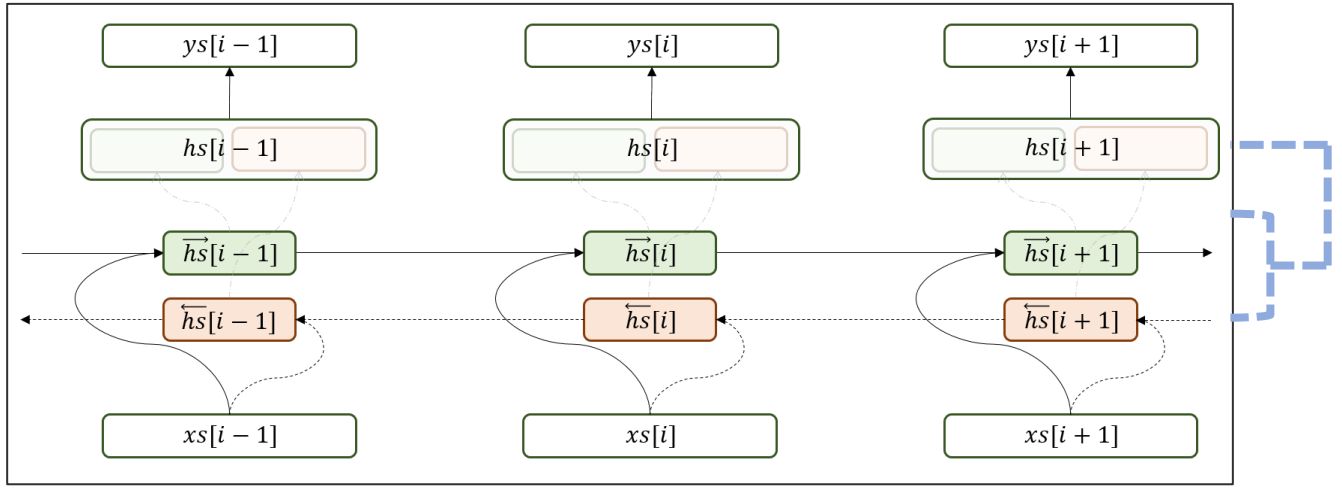
하지만 **뒤쪽을 먼저** 본다면?

P(x | ___를 뒤집어 쓰고 우는 상황과 표현) → $x$는 '이불'인 상황이 자주 발생

→ $x$ **추론 가능**

## Bidirectional GRU

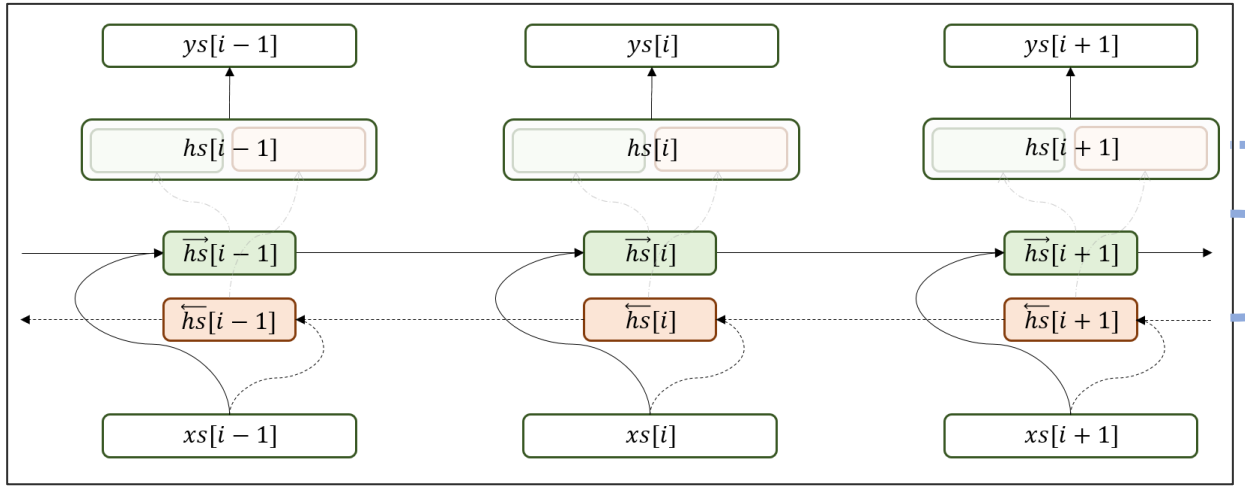
양방향(Bidirection) 재귀 모델은 이런 문제를 해결




정방향/역방향 결과값을
Concat하여 사용

## Bidirectional GRU

양방향(Bidirection) 재귀 모델은 이런 문제를 해결




정방향/역방향 결과값을 Concat하여 사용

**정방향**

앞쪽 노드일수록 정보량 적음

**역방향**

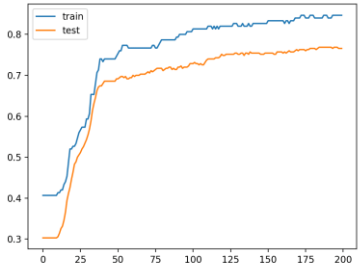
앞쪽 노드일수록 정보량 많음

각 스텝에서 **합쳐진 값들은 정보량 균등**하게 분산

# 04. 욕설탐지

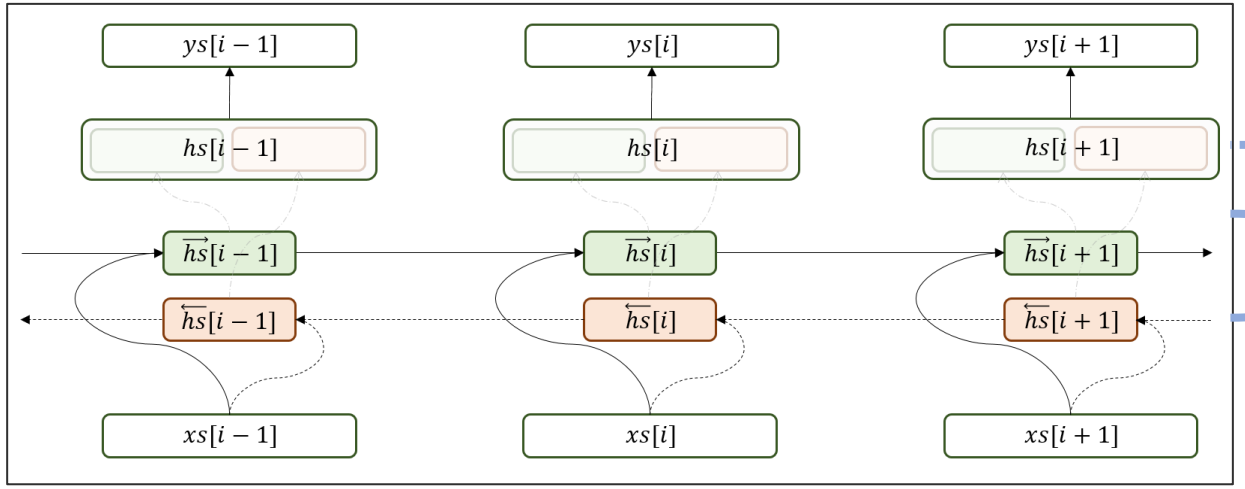
## Bidirectional GRU

욕설 데이터 학습!




Test acc: 0.7898…

양방향(Bidirection) 재귀 모델은 이런 문제를 해결




정방향/역방향 결과값을 Concat하여 사용

**정방향**

앞쪽 노드일수록 정보량 적음

**역방향**

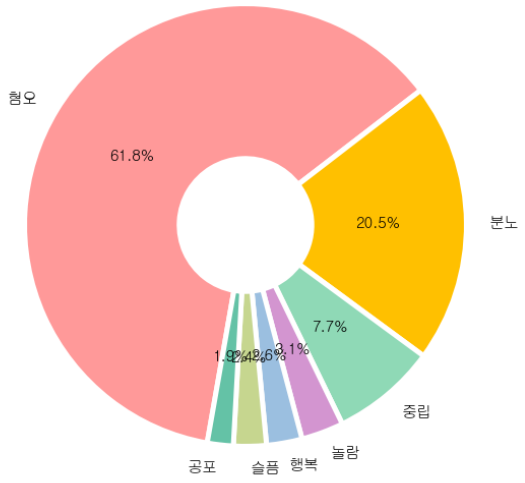
앞쪽 노드일수록 정보량 많음

각 스텝에서 **합쳐진 값들은 정보량 균등**하게 분산

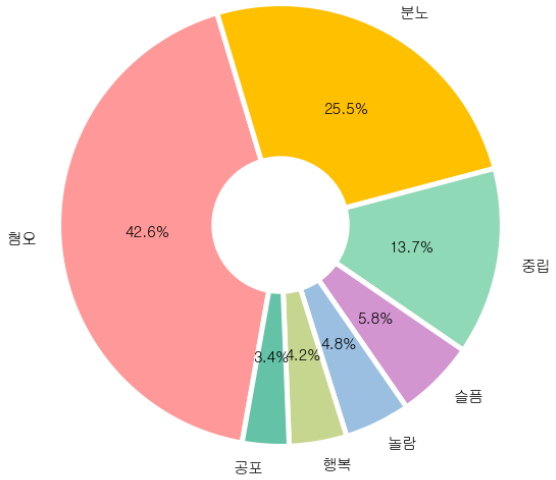
# 04. 욕설탐지

## 결과

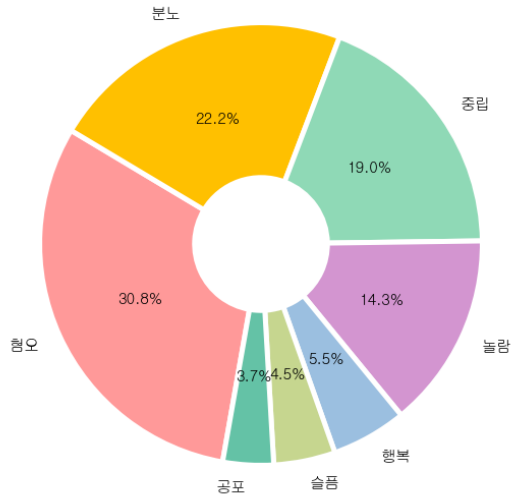
### 정치




### 경제




### 사회

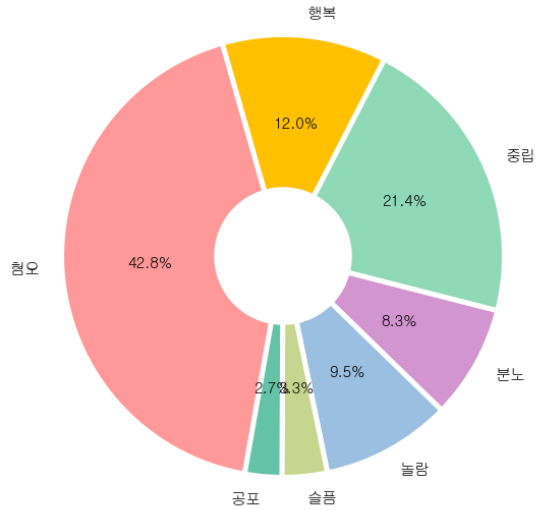



**혐오와 분노**가 많은 비율을 차지하고 있다는 것을 확인

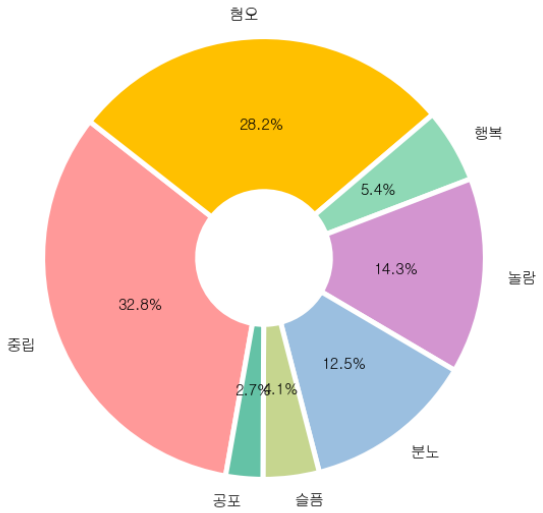
# 04. 욕설탐지

## 결과

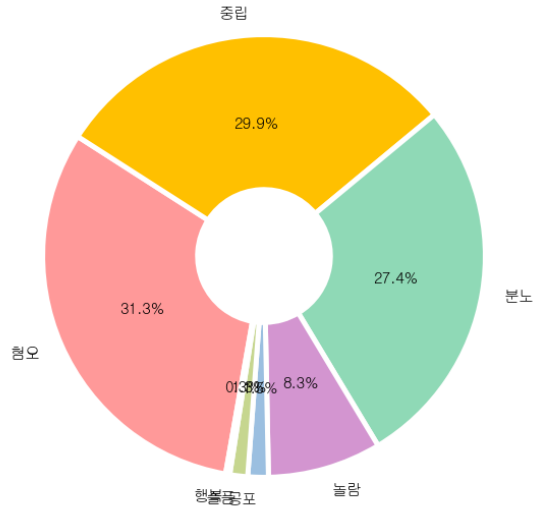
생활문화




과학




세계




**혐오와 중립**이 많은 비율을 차지하고 있다는 것을 확인

# 04. 욕설탐지

## 결과 해석

분노/혐오/중립 감정에서 많은 욕설댓글 탐지

특히 정치/경제 카테고리는 혐오+분노의 비율이 60~80%

| | |
|---|---|
| 사람 새 퀴 맞나? | 중립 |
| 여기 기사에 조선족중국인들 단체로 관광왔네. 헛소리말고 너네 연변으로 가라. 잊 | 중립 |
| 저늠도 취임식에 초대를 했었어? 거의 짐승급이던데...혹 이승만의 정치깡패 이정재 | 중립 |
| 걍 사기꾼임. 원래 사기꾼이 사기쳐놓고 이말저말 하면서 시간끌거든. | 중립 |
| 석렬이 애비 일본장학생 1호는 아떻게 생각 하는고?? | 중립 |
| 앞으로도 어쩔 수 없이 자의든 타의든 국민의 힘의 강세는 계속되겠구나. 극혐 정청 | 중립 |
| 바이든 한테 호구딜 당하는거 보더니 인니찌끄레기도 kf21호구딜 하러 오네 | 중립 |

분노/혐오 댓글 외 나머지 감정에서 욕설로 감지된 댓글 따로 확인!

카테고리별로

혐오(or 분노)의 비율 4~7%p 더욱 클 가능성 有

(특히 정치분야에서는 13% 추가탐지)

# 04. 욕설탐지

## 결과 해석

분노/혐오/중립 감정에서 많은 욕설댓글 탐지

특히 정치/경제 카테고리는 혐오+분노의 비율이 60~80%

| | |
|---|---|
| 사람 새 퀴 맞나? | 중립 |
| 여기 기사에 조선족중국인들 단체로 관광왔네. 헛소리말고 너네 연변으로 가라. 잊 | 중립 |
| 저늠도 취임식에 초대를 했었어? 거의 짐승급이던데...혹 이승만의 정치깡패 이정재 | 중립 |
| 걍 사기꾼임. 원래 사기꾼이 사기쳐놓고 이말저말 하면서 시간끌거든. | 중립 |
| 석렬이 애비 일본장학생 1호는 아떻게 생각 하는고?? | 중립 |
| 앞으로도 어쩔 수 없이 자의든 타의든 국민의 힘의 강세는 계속되겠구나. 극혐 정청 | 중립 |
| 바이든 한테 호구딜 당하는거 보더니 인니찌끄레기도 kf21호구딜 하러 오네 | 중립 |

분노/혐오 댓글 외 나머지 감정에서 욕설로 감지된 댓글 따로 확인!

카테고리별로

혐오(or 분노)의 비율 4~7%p 더욱 클 가능성 有

(특히 정치분야에서는 13% 추가탐지)


# 5. LDA

## LDA (Latent Dirichlet Allocation)

그럼 이제 감정별로

**어떤 토픽**이 나타나는지 LDA로 분석해보자

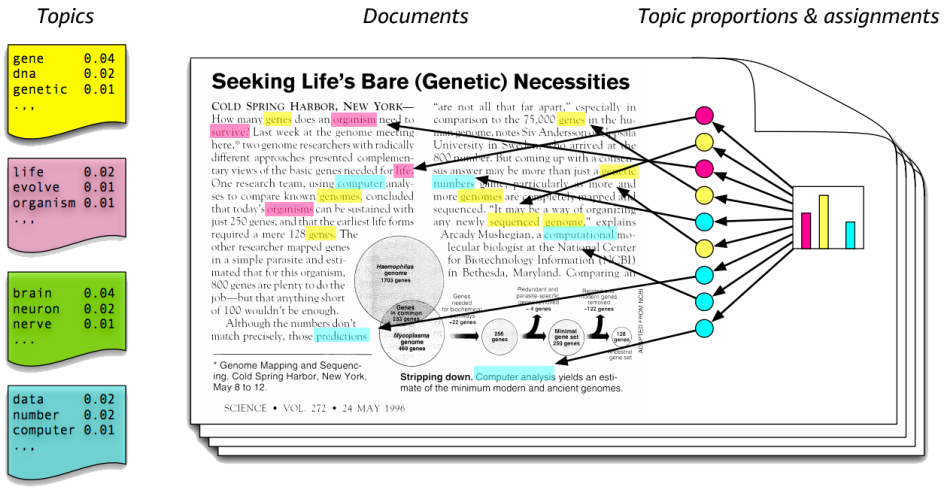
# LDA (Latent Dirichlet Allocation)

## LDA (Latent Dirichlet Allocation)

확률분포 중 하나인 디리클레 분포를 가정

번호가 매겨진 토픽 안에 문서와 단어들을 하나씩 넣어보며 잠재적인 토픽들을 찾아주는 과정

*디리클레 분포: 베타 분포를 k가지 경우를 다루도록 확정*




**베이즈 추론**에 기반

$$P(\theta|X) = \frac{p(\theta, X)}{p(X)} = \frac{p(X|\theta)p(\theta)}{p(X)}$$

1. 어떤 사건이 발생할 확률 가정
2. 추가적인 관측 발생 시, 그 사건이 발생할 확률을 더 정확하게 추론

## 최적의 파라미터

### 최적의 파라미터의 기준

#### Perplexity

특정 확률 모델이 실제로 관측되는 값을
얼마나 잘 예측하는지

**작을수록** 토픽모델이 문서를 잘 반영함

#### Coherence

주제의 일관성
토픽이 얼마나 의미론적으로 일관성 있는지

**높을수록** 일관성 높음

## 최적의 파라미터

최적의 파라미터의 기준

Perplexity

혼란도

특정 확률 모델이 실제로 관측되는 값을

얼마나 잘 예측하는지

**작을수록** 토픽모델이 문서를 잘 반영함

Coherence

주제의 일관성

토픽이 얼마나 의미론적으로 일관성 있는지

**높을수록** 일관성 높음

Coherence가 너무 **높으면 정보의 양이 줄어들고**,

Coherence가 낮아 정보들의 인과성이 없다면, 분석의 의미가 없어짐

Bias-Variance Tradeoff처럼 **적당한 합의점**을 찾는 것이 중요

# Grid Search

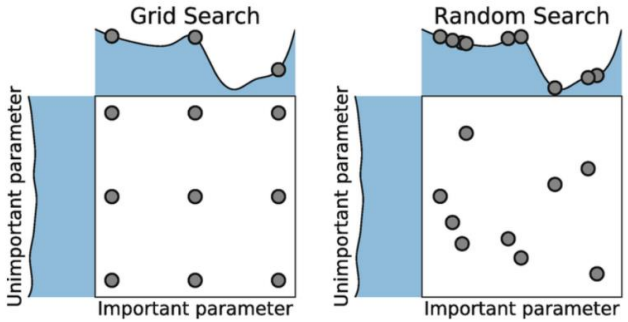
### 하이퍼 파라미터 튜닝이란?

모델의 성능을 확보하기 위해 주요 설정 값 조정

### 그리드 서치 (Grid Search)

하이퍼 파라미터 후보들을 **하나씩 입력**

모델의 성능을 **가장 높게** 하는 최적의 파라미터를 찾음




Perplexity 값이 **최소**가 되도록

Coherence가 **최대**가 되도록

토픽 개수와 learning rate 조절

## Grid Search

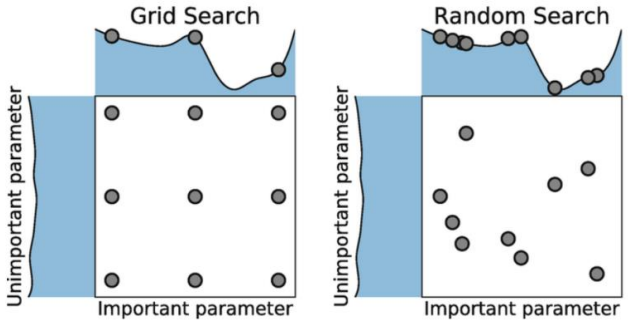
### 하이퍼 파라미터 튜닝이란?

모델의 성능을 확보하기 위해 주요 설정 값 조정

### 그리드 서치 (Grid Search)

하이퍼 파라미터 후보들을 **하나씩 입력**

모델의 성능을 **가장 높게** 하는 최적의 파라미터를 찾음




Perplexity 값이 **최소**가 되도록

Coherence가 **최대**가 되도록

토픽 개수와 $\alpha, \beta$조절

## 하이퍼 파라미터 튜닝

학습 파라미터

1. Topic : 토픽의 개수
2. Chunksize: 알고리즘에 사용된 문서 수
3. Passes : 전체 코퍼스에서 모델을 학습시키는 빈도 (epoch)
4. iteration : 각각 문서에 대해서 루프를 얼마나 돌리는지
5. Alpha, Beta: 디리클레 분포의 감마함수에 대한 파라미터

$$p(\theta, z, w \mid \alpha, \beta) = p(\theta \mid \alpha) \prod_{n=1}^{N} p(z_n \mid \theta) p(w_n \mid z_n, \beta)$$



coherence

Umass score

두 단어가 같이 등장할 확률

$D(w_i, w_j)$: $w_i$와 $w_j$가 같이 등장할 확률

$D(w_i)$: $w_i$가 등장할 확률

$$C_{U\,Mass}(w_i, w_j) = \log \frac{D(w_i, w_j) + 1}{D(w_i)}$$

## 하이퍼 파라미터 튜닝



학습 파라미터

1. Topic : 토픽의 개수
2. Chunksize: 알고리즘에 사용된 문서 수
3. Passes : 전체 코퍼스에서 모델을 학습시키는 빈도 (epoch)
4. iteration : 각각 문서에 대해서 루프를 얼마나 돌리는지
5. Alpha, Beta: 디리클레 분포의 감마함수에 대한 파라미터

coherence

Umass score

두 단어가 같이 등장할 확률

$P(w_i, w_j)$: $w_i$와 $w_j$가 같이 등장할 확률

$P(w_i)$: $w_i$가 등장할 확률

$\epsilon$ : 1보다 작은 값

$$C_{U\,Mass} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{P(w_i, w_j) + \epsilon}{P(w_j)}$$

*출처: Exploring the Space of Topic Coherence Measures*

## 전처리

### 1. 각 분야를 감정별로 쪼개기

| comment | emotion |
|---|---|
| 정동원 선한영향력 모범청소년 멋지게성장중인 만능엔터테이너 응원합니다 | 행복 |
| 뉴진스의 인기를 보며 나이가 들었다는게 확 느껴진다 | 중립 |
| 포레스트검프 부터 보고 봐야지 | 중립 |
| 칼로 물베기...왜하냐...뭘 기대하는지 | 분노 |

| comment | emotion |
|---|---|
| 뉴진스의 인기를 보며 나이가 들었다는게 확 느껴진다 | 중립 |
| 포레스트검프 부터 보고 봐야지 | 중립 |
| 두방이냐 세방이냐가 관건 | 중립 |
| 문제생길줄 알았다 | 중립 |

## 전처리

2. 글자, 숫자 추출 ▪▪▪▶ 3. 명사 추출 ▪▪▪▶ 4. 토큰화

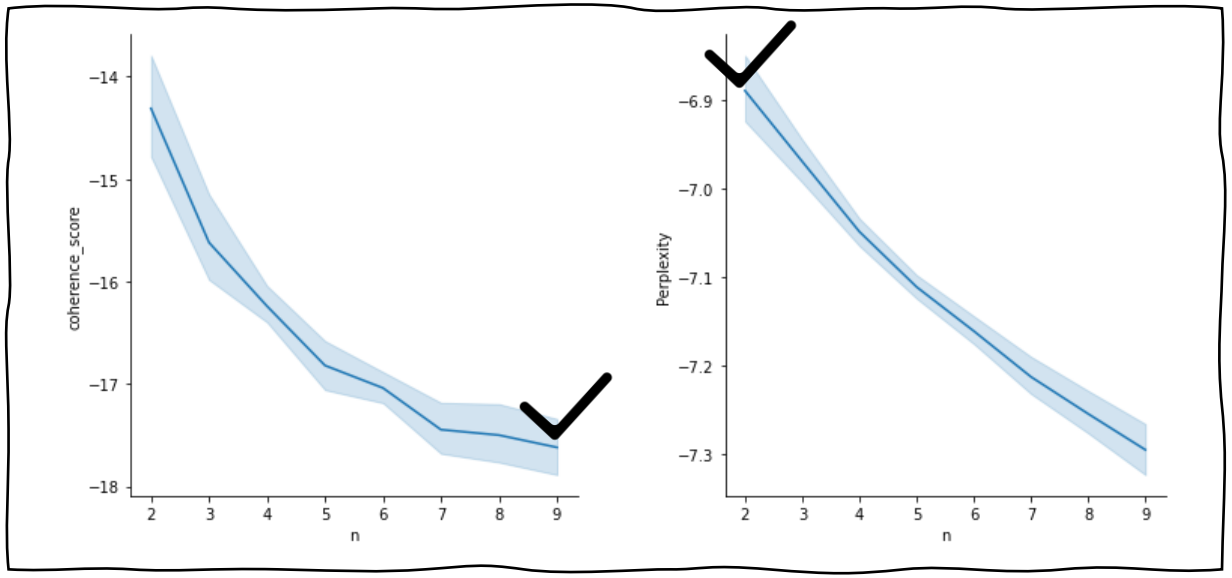
['진스', '인기', '나이'],
['포레스트', '검프'],
['옛날', '건물', '당연'],
['크루즈', '한국', '사연', '궁금', '지네'],
['예전', '공주', '토끼', '기억', '그때', '맛있']

## 전처리

### 5. 토픽 모델링

Grid search로 coherence score와 perplexity 튜닝

최적의 결과를 선정 후 model fit!




Perplexity는 negative log perplexity로 출력

→ 값이 클수록 좋음

Coherence는 U_Mass 사용

→ 작을수록 좋음


## 전처리

### Q1) 앞에서 Perplexity는 작을수록 좋다고 하지 않았나요…?

작은 값에 로그와 마이너스를 붙이면 값이 커짐

$-\log(1.3) \rightarrow -0.11394$ / $-\log(1.7) \rightarrow -0.23044$

로그를 제외한 실제 값을 비교하면 비록 1.3이 1.7보다 작지만 마이너스 로그를 적용하면 반대가 됨

### Q2) 앞에서 Coherence는 높을수록 좋다고 하지 않았나요…?

$$C_{U\ Mass} = \frac{2}{N(N-1)} \sum_{i=2}^{N} \sum_{j=1}^{i-1} log \frac{P(w_i, w_j) + \epsilon}{P(w_j)}$$

확률 값을 입력하기 때문에 분자 < 분모

위와 같은 맥락

0이 나오면 완벽하다고 판단

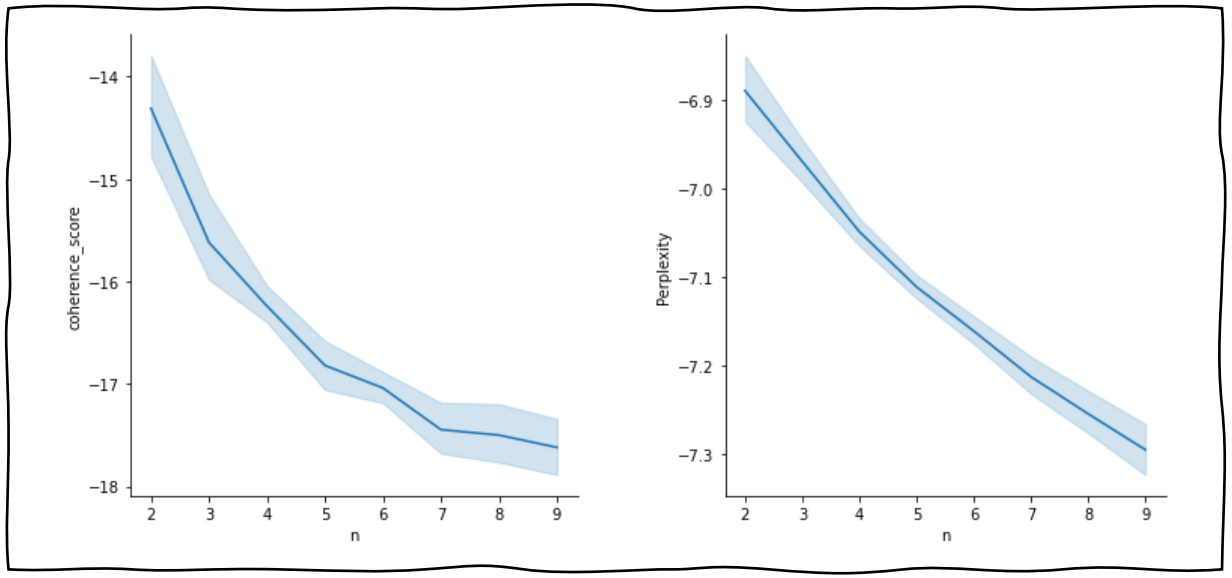
→ 작을수록 좋음

## 전처리

### 5. 토픽 모델링

Grid search로 coherence score와 perplexity 튜닝

최적의 결과를 선정 후 model fit!




Perplexity와 Coherence의 값들이

크게 차이 나지 않는 것을 고려,

Heuristic 하게 정함

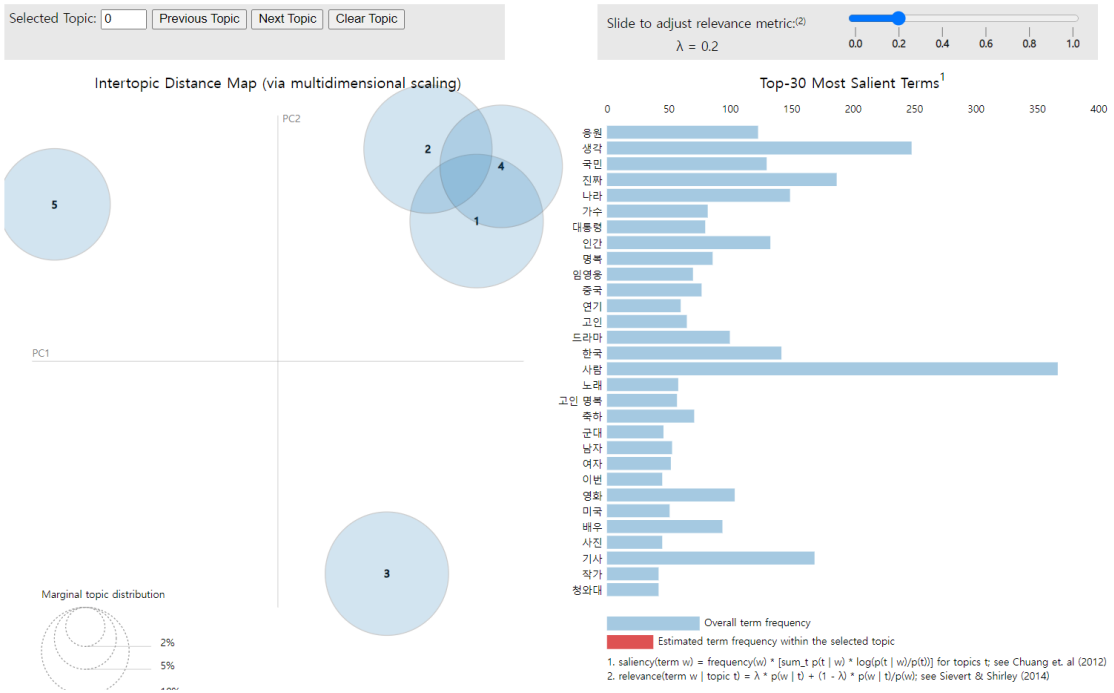
n = 3

# LDAvis

## LDAvis

LDA 결과를 시각적으로 표현하는 라이브러리

EXAMPLE) 생활문화

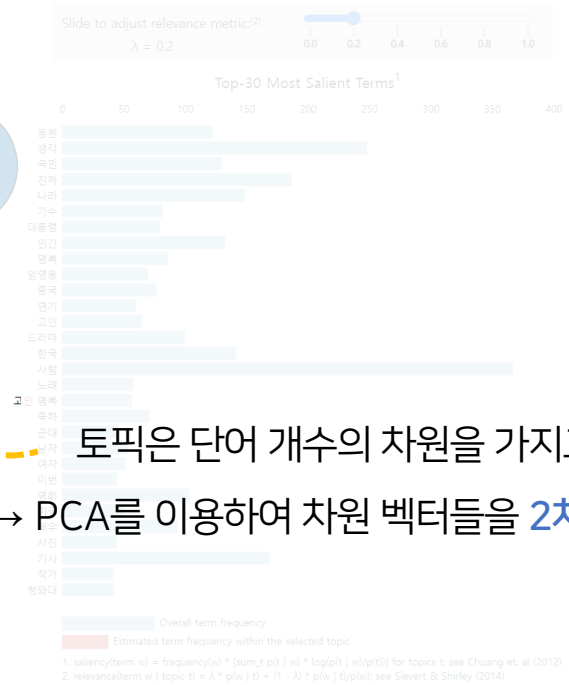


# LDAvis

## LDAvis

LDA 결과를 시각적으로 표현하는 라이브러리

EXAMPLE) 생활문화


**원의 크기**: 토픽의 단어들이 얼마나 속해 있고 어떻게 분포되어 있는지

**원의 거리**: 토픽 간의 유사성

토픽은 단어 개수의 차원을 가지고 있음

→ PCA를 이용하여 차원 벡터들을 **2차원**으로 압축

# LDAvis

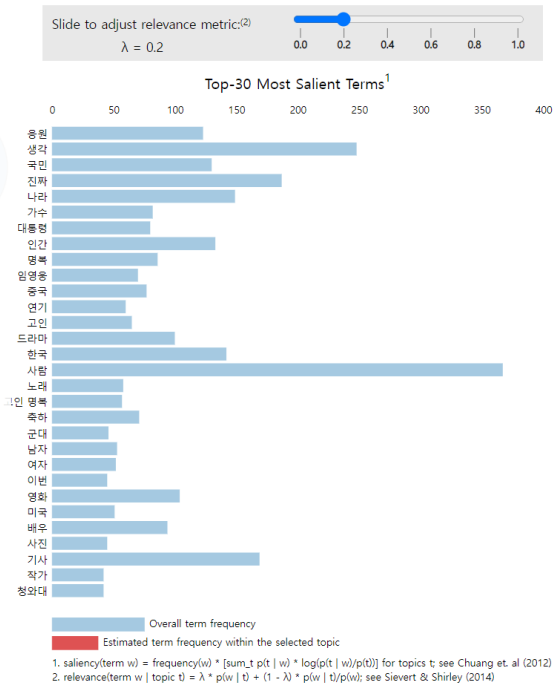
LDAvis

LDA 결과를 시각적으로 표현하는 라이브러리

Top-30 Most Relevant Terms for Topic

토픽을 형성하는 **주도적인 키워드** 확인

(각 키워드마다 토픽과의 관련성 확인 가능)



## LDAvis

Keyword Extraction 기준

### Salience P(w|t)

단어 w를 갖고 있는 모든 문서들 중

**토픽 t가 할당된 비율**

문제점: 특정 토픽을 나타내지 않는 차별성이 없는

단어는 P(w|t)가 높음

### Discriminative Power P(w|t)/P(w)

각 토픽에서의 단어 발생 확률을

단어의 기본 발생 확률로 **정규화**

문제점: 한 토픽에서만 등장한 단어는

전체에서도 많이 등장하지 않을 확률이 높음

## LDAvis

Keyword Extraction 기준

### Salience P(w|t)

단어 w를 갖고 있는 모든 문서들 중
토픽 t가 할당된 비율

문제점: 특정 토픽을 나타내지 않는 차별성이 없는
단어는 P(w|t)가 높음

### Discriminative Power P(w|t)/P(w)

각 토픽에서의 단어 발생 확률을
단어의 기본 발생 확률로 정규화

문제점: 한 토픽에서만 등장한 단어는
전체에서도 많이 등장하지 않을 확률이 높음

Negative correlation 관계

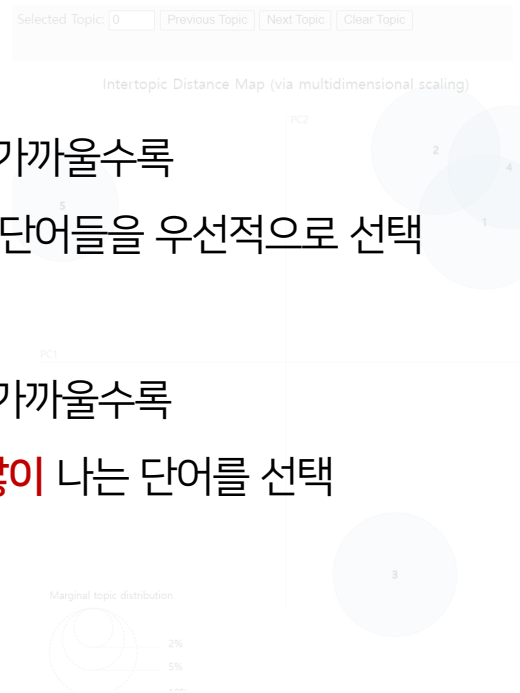
**양면을 모두 고려**하여 키워드 선택!

# LDAvis

LDAvis

LDA 결과를 시각적으로 표현하는 라이브러리

EXAMPLE) 생활문화

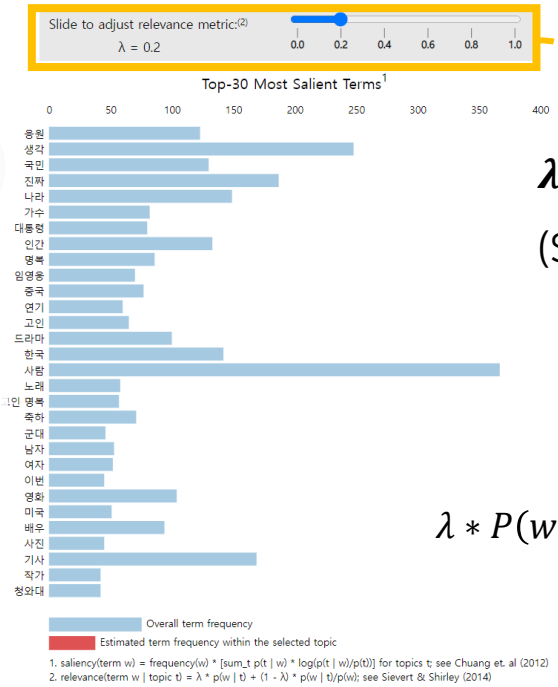



$\lambda$가 1에 가까울수록

토픽별로 **자주 등장하는** 단어들을 우선적으로 선택

$\lambda$가 0에 가까울수록

토픽 간에 **차이가 많이** 나는 단어를 선택




**$\lambda$값**

(Slide to adjust relevance metric)

키워드 랭킹 점수 계산:

$$\lambda * P(w|t) + (1-\lambda) * P(w|t)P(w)\lambda * P(w|t)$$

$$+(1-\lambda) * P(w|t)p(w)$$

## LDAvis

감정별로 토픽이 어떻게 나뉘는지

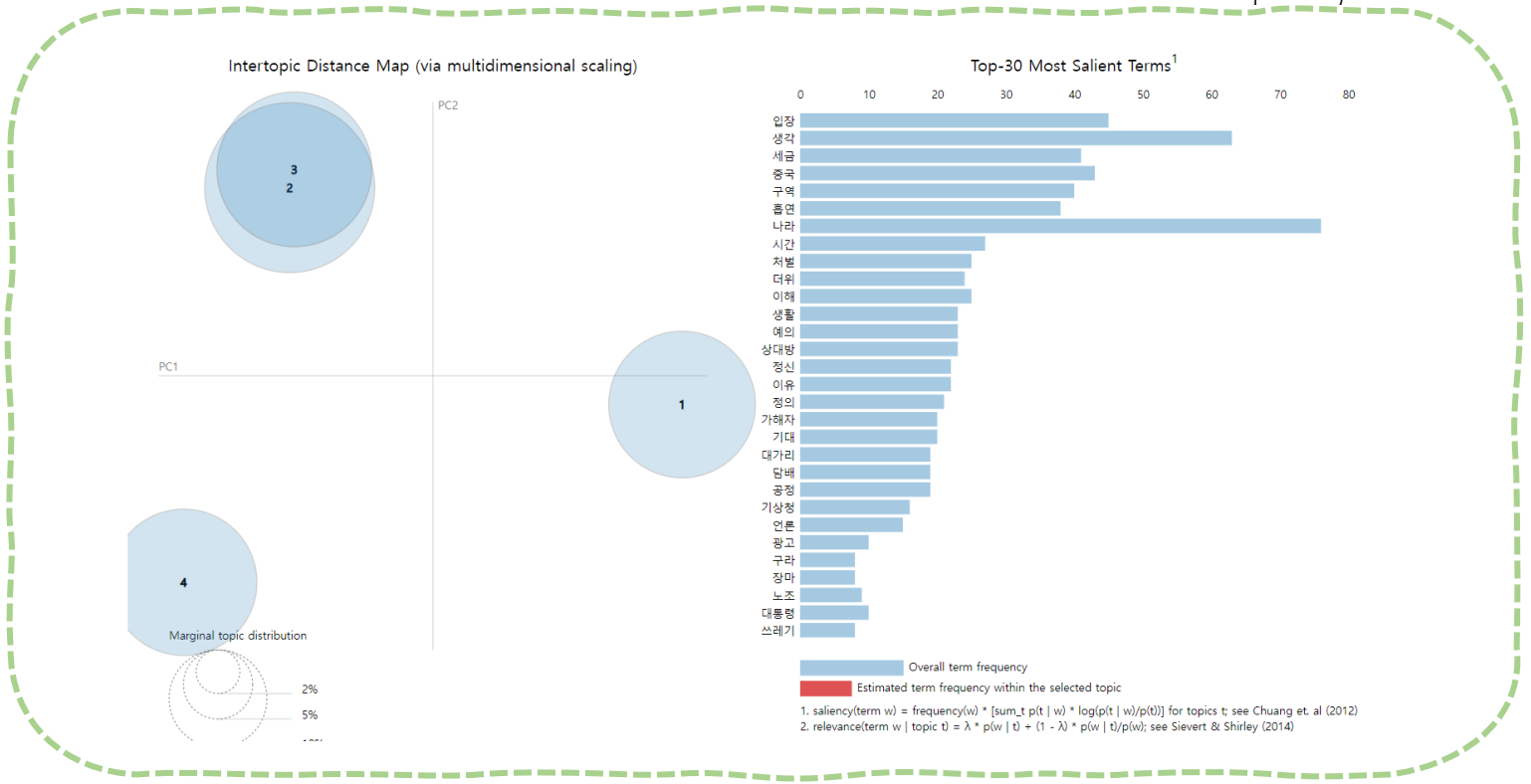
그 **차이점**에 대해서 알아보는 게  목적!

$\lambda$값을 0.5에서 줄여가면서

키워드들이 과도하게 세분화되지 않는 지점에서 분석

# 05. LDA

## 결과 해석

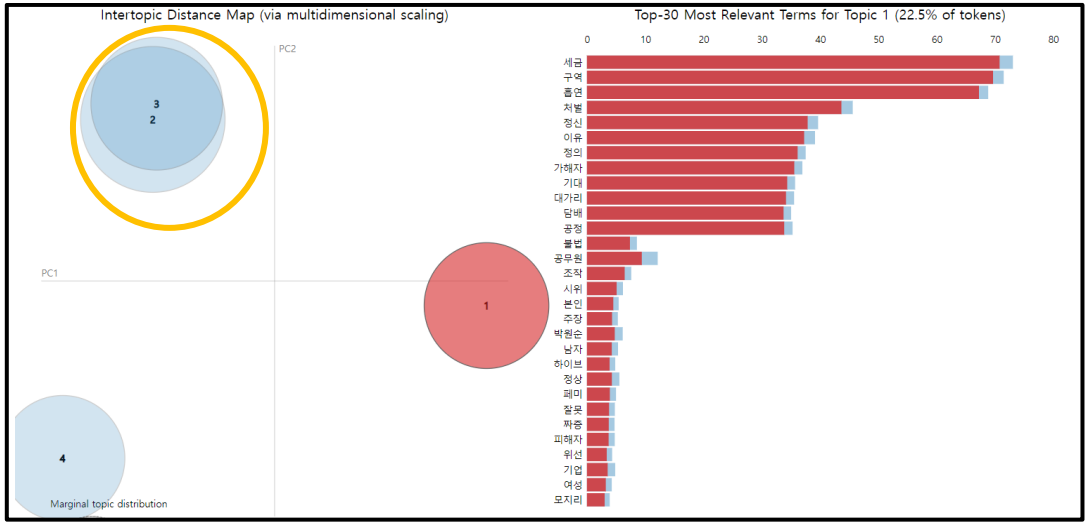



생활문화 – 혐오 데이터 → **4가지** 토픽

## 결과 해석

생활문화 - 혐오

EXAMPLE) 1번 토픽




토픽 내 중요 키워드

| Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|
| 세금 | 나라 | 중국 | 더위 |
| 구역 | 대통령 | 언론 | 생활 |
| 흡연 | 공개 | 광고 | 예의 |
| 처벌 | 문재인 | 구라 | 쓰레기 |
| 정신 | 인간 | 기사 | 동물 |

유사성 높음

2~30이 겹쳐 있음 → **유사성 높음**

## 결과 해석

생활문화 - 혐오

| Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|
| 세금 | 나라 | 중국 | 더위 |
| 구역 | 대통령 | 언론 | 생활 |
| 흡연 | 공개 | 광고 | 예의 |
| 처벌 | 문재인 | 구라 | 쓰레기 |
| 정신 | 인간 | 기사 | 동물 |

각 토픽별로 댓글(토픽 추출 과정에서 끼친 영향)

**중요도 순**으로 확인

Topic1----------

왜 증여나 상속을 비난하나. 자식 고생시키지 않으려고 온갖 고생하면서 돈 버는 거다/문제는 법이 정한 절차를 밟아 정당하게…

Topic2----------

일본이 코로나 때문에 참석이 저조하다고 외람되게 기사를 쓸 수 있냐 이말입니다/터키가 왜??우리의 형제국인가요?? 전 아니라고 생각…

Topic3----------

서울 경제 기자양반 하나만 물어 봅시다 그쪽 동네 기자들은 문재인 정부 방역이 X같다고 대놓고 대한민국이 망했다고 하던데…

Topic4----------

어차피 용기에 섭취방법과 주의사항이 있는데 뭐가 문제라는건지...

그냥 쓰레기 시민,관변 단체들이 설쳐대니 규제만 만들었는건…

## 결과 해석

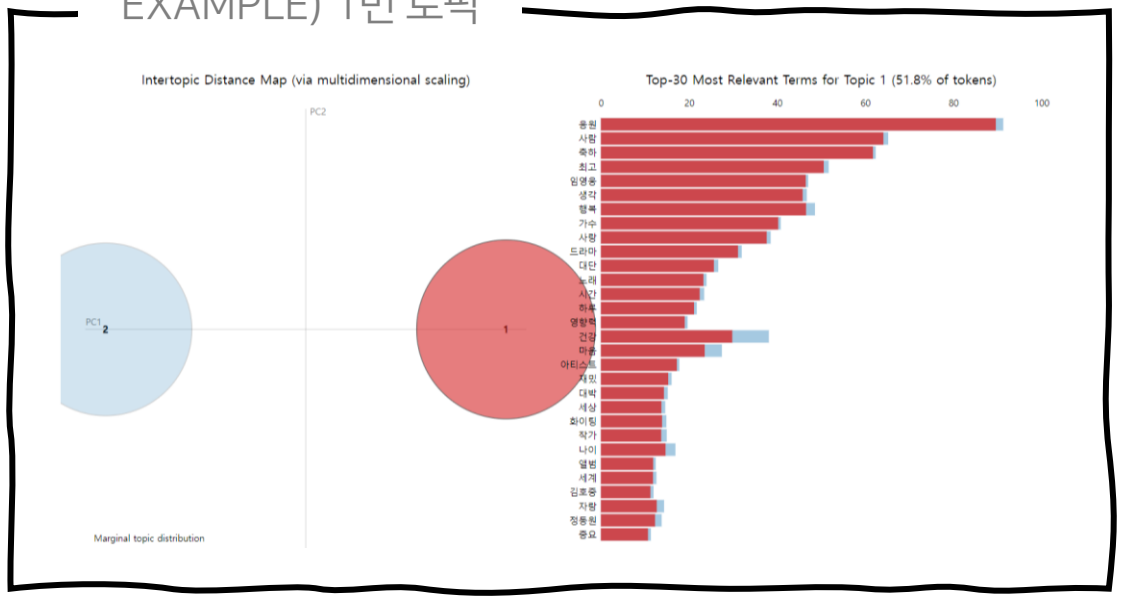
비판적인 시각

- 법/처벌
  - Topic1----------
  - 왜 증여나 상속을 비난하나. 자식 고생시키지 않으려고 온갖 고생하면서
  - 돈 버는 거다/문제는 법이 정한 절차를 밟아 정당하게…
- 정부/외교
  - Topic2----------
  - 일본이 코로나 때문에 참석이 저조하다고 외람되게 기사를 쓸 수 있냐
  - 이말입니다/터키가 왜??우리의 형제국인가요?? 전 아니라고 생각…
- 언론/외교
  - Topic3----------
  - 서울 경제 기자양반 하나만 물어 봅시다 그쪽 동네 기자들은 문재인 정부
  - 방역이 X같다고 대놓고 대한민국이 망했다고 하던데…
- 기상청/생활
  - Topic4----------
  - 어차피 용기에 섭취방법과 주의사항이 있는데 뭐가 문제라는건지...
  - 그냥 쓰레기 시민,관변 단체들이 설쳐대니 규제만 만들었는건…

## 결과 해석

생활문화-행복

EXAMPLE) 1번 토픽




토픽 내 중요 키워드

| Topic 1 | Topic 2 |
|---|---|
| 응원 | 감사 |
| 사람 | 배우 |
| 축하 | 기대 |
| 최고 | 영화 |
| 임영웅 | 연기 |

2가지 토픽 → 댓글 **분포/개수 유사**(509:470)

## 결과 해석

생활문화-**행복**

| Topic 1 | Topic 2 |
|---|---|
| 응원 | 감사 |
| 사람 | 배우 |
| 축하 | 기대 |
| 최고 | 영화 |
| 임영웅 | 연기 |

각 토픽별로 댓글(토픽 추출 과정에서 끼친 영향)

**중요도 순**으로 확인

Topic1----------

송해 선생님께서 전국 노래자랑 그 자체셨는데 상상이 되지를 않는다. 건강에 이상이있으셔서 그만 두시는게 아니길 바라고 송해 선생님께서 건강하셨으면 한다. 그동안 고생 정말 많으셨습니다./인성 갑 노래 갑 임영웅가수님 언제나 응원합니다./방탄소년단의 앞으로의 활동도 기대되고 항상 응원합니다 이번 앨범 너무 좋아요~/연주할때 모습이 얼마나 …

Topic2----------

이 드라마는 어디까지나 드라마를 위한,드라마 이기때문에,드라마답게 자폐에대한 미화가 있는 내용임에는 분명하다. 그러나 이 드라마를 보고 마음이 따듯해짐을 느끼듯이 정말 현실속 모든 장애인에대한 따듯한 시선과 배려를 갖을수 있게되길 바래봅니다./오래 전 커피프린스 1호점…

## 결과 해석


긍정적 시각

- 음악
  - Topic1----------
    송해 선생님께서 전국 노래자랑 그 자체셨는데 상상이 되지를 않는다. 건강에 이상이있으셔서 그만 두시는게 아니길 바라고 송해 선생님께서 건강하셨으면 한다. 그동안 고생 정말 많으셨습니다./인성 갑 노래 갑 임영웅가수님 언제나 응원합니다./방탄소년단의 앞으로의 활동도 기대되고 항상 응원합니다 이번 앨범 너무 좋아요~/연주할때 모습이 얼마나 …
- 드라마/영화
  - Topic2----------
    이 드라마는 어디까지나 드라마를 위한,드라마 이기때문에,드라마답게 자폐에대한 미화가 있는 내용임에는 분명하다. 그러나 이 드라마를 보고 마음이 따듯해짐을 느끼듯이 정말 현실속 모든 장애인에대한 따듯한 시선과 배려를 갖을수 있게되길 바래봅니다./오래 전 커피프린스 1호점…

## 결과 해석

생활문화 - 분노

| Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|
| 표절 | 드라마 | 기자 | 세금 |
| 불편 | 기분 | 한심 | 중국 |
| 부담 | 댓글 | 남자 | 정부 |
| 국민 | 해외 | 자식 | 연설문 |
| 생각 | 감독 | 부모 | 박근혜 |

각 토픽별로 댓글(토픽 추출 과정에서 끼친 영향)

중요도 순으로 확인

Topic1----------

노인보호센타 아무런 도움이 안된다/사법권이 없어서... 일을 더진행 할 수 없다면서... 신고한 노인더러 본인이 경찰에 고소를 하라고 하면서... 자기들은 빠진다. 아무런 도움이 안된다./이웃은 악이든 선이든 상관…

Topic2----------

영화줄거리나 관전평에 대해 본적도 들은적도 없는데 사이비 좌빨 양아치 언론 한 걸레가 저런식으로 음모론까지 강조해 기사 쓴걸보니 믿고…

Topic3----------

◆ 무식한 윤석열 새긔. 돼지보다 더 처먹는거 좋아하던데 ㅉㅉ 바이든과 만날 때 급똥 마려워서 망신당하길 기원합니다 ㅉㅉ 바이든은 …

Topic4----------

정치진영을 떠나서 친중은 정말 어리석어요...친중정책 정당은 배척이 맞습니다! 중국에서 살아보니 확실하고 확신합니다…

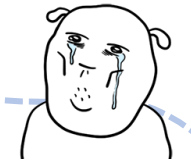
## 결과 해석

생활문화 - 슬픔

| Topic 1 | Topic 2 | Topic 3 | Topic 4 |
|---|---|---|---|
| 고생 | 교육 | 명복 | 우울증 |
| 문제 | 부럽 | 고인 | 생계 |
| 미안 | 거지 | 영화 | 불쌍 |
| 사회 | 환경 | 배우 | 후회 |
| 피해 | 물가 | 우상 | 사망 |

각 토픽별로 댓글(토픽 추출 과정에서 끼친 영향)

중요도 순으로 확인



Topic1----------

애들보다보면 하루가 지친다 씻기고 밥해먹이고 가정주부도 하루가 지친다 숙제 공부봐주고 밤되면 피곤해서 곯아떨어진다 이런일들이 365일 반복하며 산다 너무 피곤해서 자기바쁘다…

Topic2----------

혼자서 해먹는건 밖에서 사먹는거보단 여전히 비싼데 외식값이 장난이 아닌지 슬픈 세상/믿고먹는 스벅이었는데 이제 안가요 프라푸치…

Topic3----------

안타까운 죽음이다...고인의 명복을 빕니다…/안타깝습니다. 좋은 배우셨습니다..미인박명 인지 안타깝고 애석하네요!일찍발견되 수술을…

Topic4----------

진짜 아프면 방송에 나올 힘도 없다 - 조용히 치료 받길/그래 ADHD라고 자기 위안 하고 살어./너무 안타깝다 ㅠㅠㅠ 아이치료할때는 제발…

## 결과 해석

생활문화 - 공포

| Topic 1 | Topic 2 |
|---|---|
| 재앙 | 문제 |
| 신고 | 아이 |
| 전화 | 견주 |
| 의사 | 부작용 |
| 회복 | 걱정 |

각 토픽별로 댓글(토픽 추출 과정에서 끼친 영향)

중요도 순으로 확인

Topic1----------

아직까진 괜찮지 50살 넘어서 부모 다 돌아가시고 형제자매들은 다 자기가족 챙기기 바쁜상황이 오면 돈이 많아도 뭐할래? 맹장터져도 혼자 119에 전화해서 보호자도 없이 혼자 아파봐야 정신차린다/윤재앙 핵투하/가뜩이나 경기회복으로 많은 오락거리가 살아나기 시작했고, 콘서트도 많이 했다. 하지만…

Topic2----------

저 얘를 그냥 자숙시키고 다시 내보낼 생각이라면 이건 어른들도 소름돋게 무섭다.. 피해자는 눈꼽만큼도 생각안하고 자기들만 생각하는 이기심.. 5호처분을 받을만큼 뭔가 있었다면 탈퇴정도는 감수해야하는거 아닌가 어렸을때 그냥 말한마디로도 상처받아 그게 평생을 가기도 하는데 말몇마디로 쉴드칠 일은 아니죠 /사람 물어뜯는 견의 귀여운 사진과 기죽…

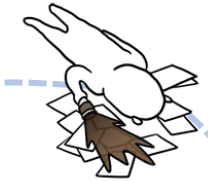
## 결과 해석

생활문화 - 놀람

| Topic 1 | Topic 2 | Topic 3 |
|---|---|---|
| 대통령 | 인기 | 영화 |
| 국민 | 마케팅 | 촬영 |
| 우리나라 | 프로 | 손흥민 |
| 정치 | 운동 | 기자 |
| 지역 | 가격 | 옥주현 |

각 토픽별로 댓글(토픽 추출 과정에서 끼친 영향)

**중요도 순**으로 확인



Topic1----------

애들보다보면 하루가 지친다 씻기고 밥해먹이고 가정주부도 하루가 지친다 숙제 공부봐주고 밤되면 피곤해서 곪아떨어진다 이런일들이 365일 반복하며 산다 너무 피곤해서 자기바쁘다…

Topic2----------

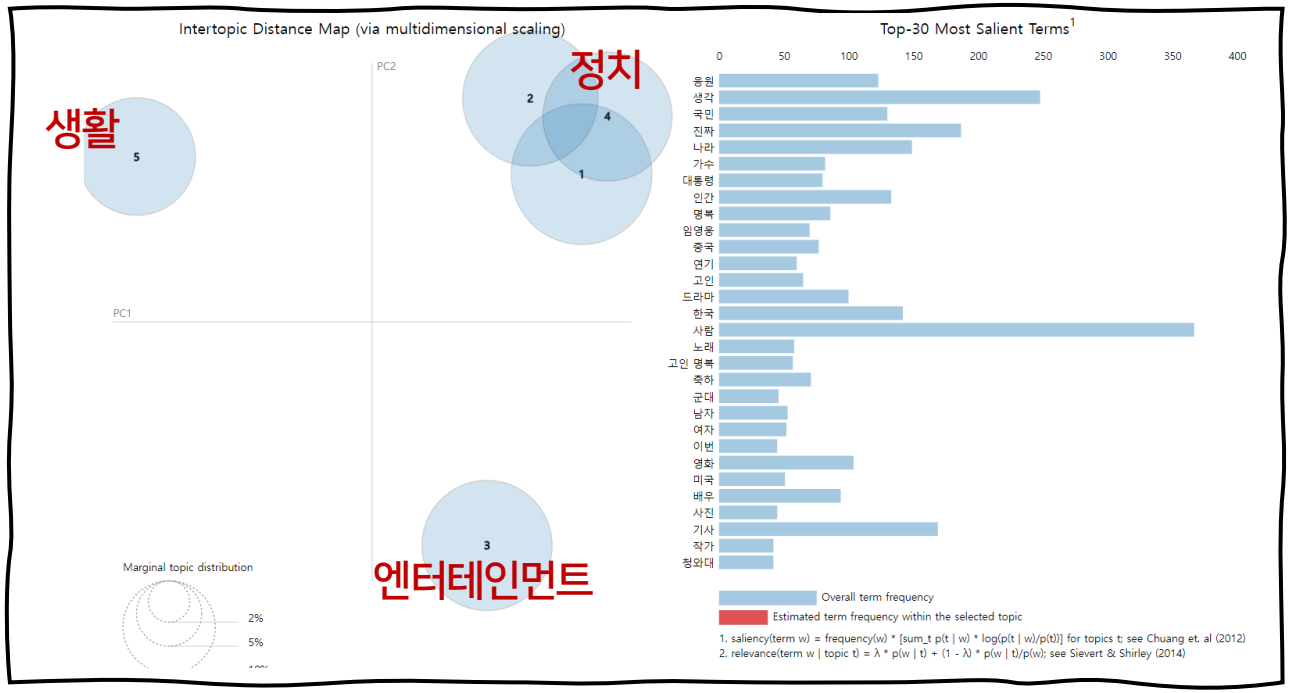
시금치가 ..6000원 아욱이 5000원...계란이8000원 니미럴/너네 20~40%편의점 수수료는?…/저건 13만원 짜린 아니지../와 싸이 이래서 복귀했구나 축제 하는지 알고 댄단한데

Topic3----------

B급전문 배우인데 대가리는 f급 상태네/노메이크업 사진은 경악 그자체였다/설사 인맥 캐스팅이라고 해도 제작사가 그렇게 하겠다는데 이게 왜?.../아이유는 왜 뮤직뱅크거 아니라 칸에 가있는거?

## 결과 해석

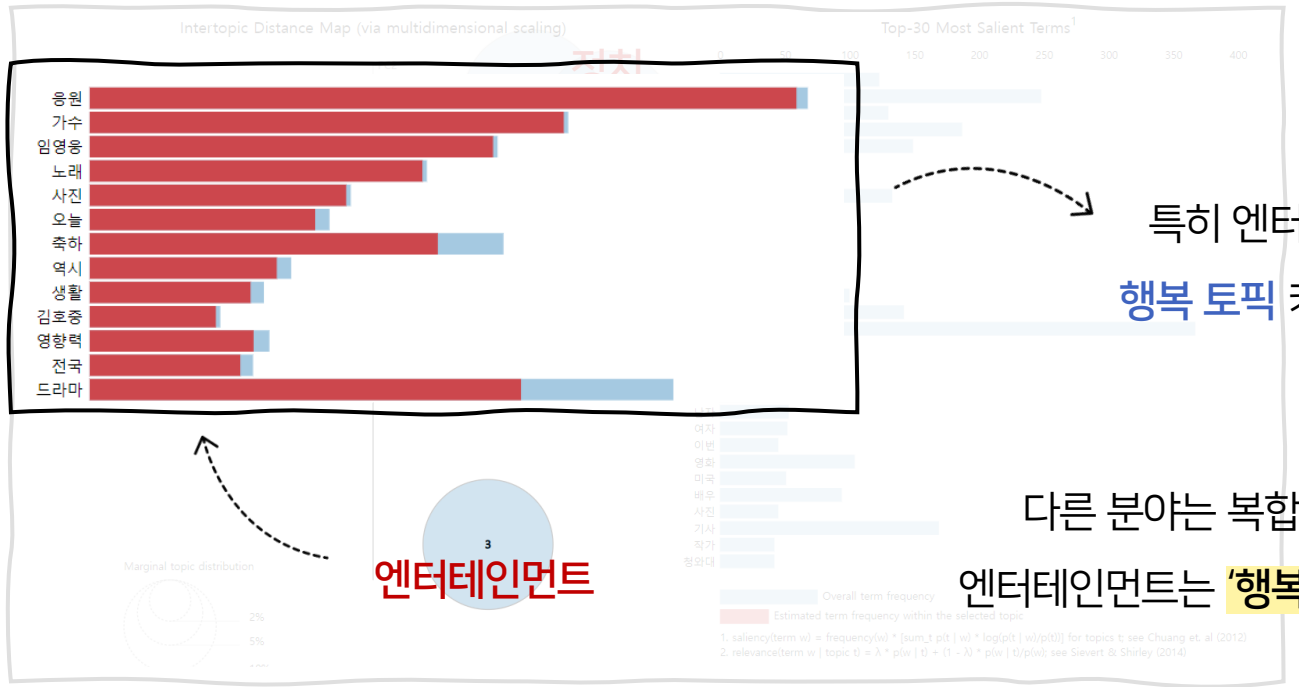
생활문화 전체




**감정구분 없이** 생활문화 전체적으로 토픽을 분석해봤을 때,

크게 세 분야(정치, 엔터테인먼트, 생활)로 구분

## 결과 해석




특히 엔터테인먼트 분야는

**행복 토픽** 키워드와 매우 유사

다른 분야는 복합적인 감정이 나타난 반면

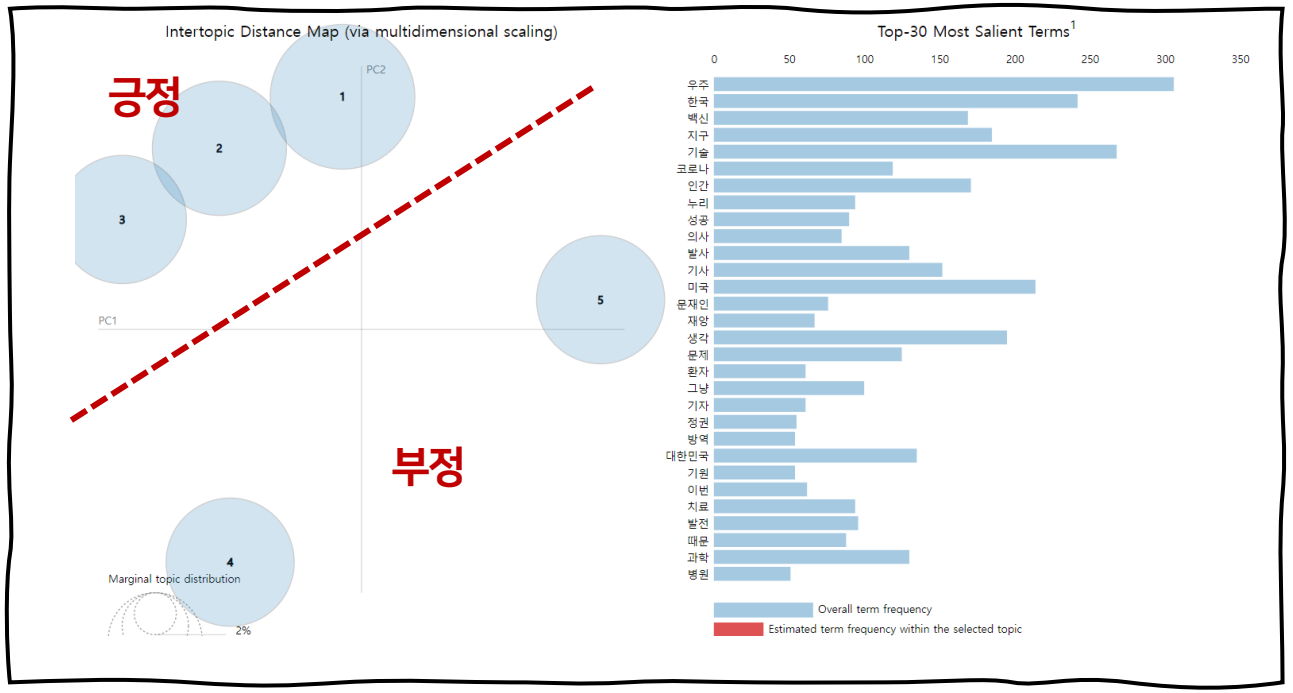
엔터테인먼트는 **'행복'이 우세**한 것으로 해석 가능

**감정구분 없이** 생활문화 전체적으로 토픽을 분석해봤을 때,

크게 세 분야(정치, 엔터테인먼트, 생활)로 구분

## 결과 해석

과학 전체

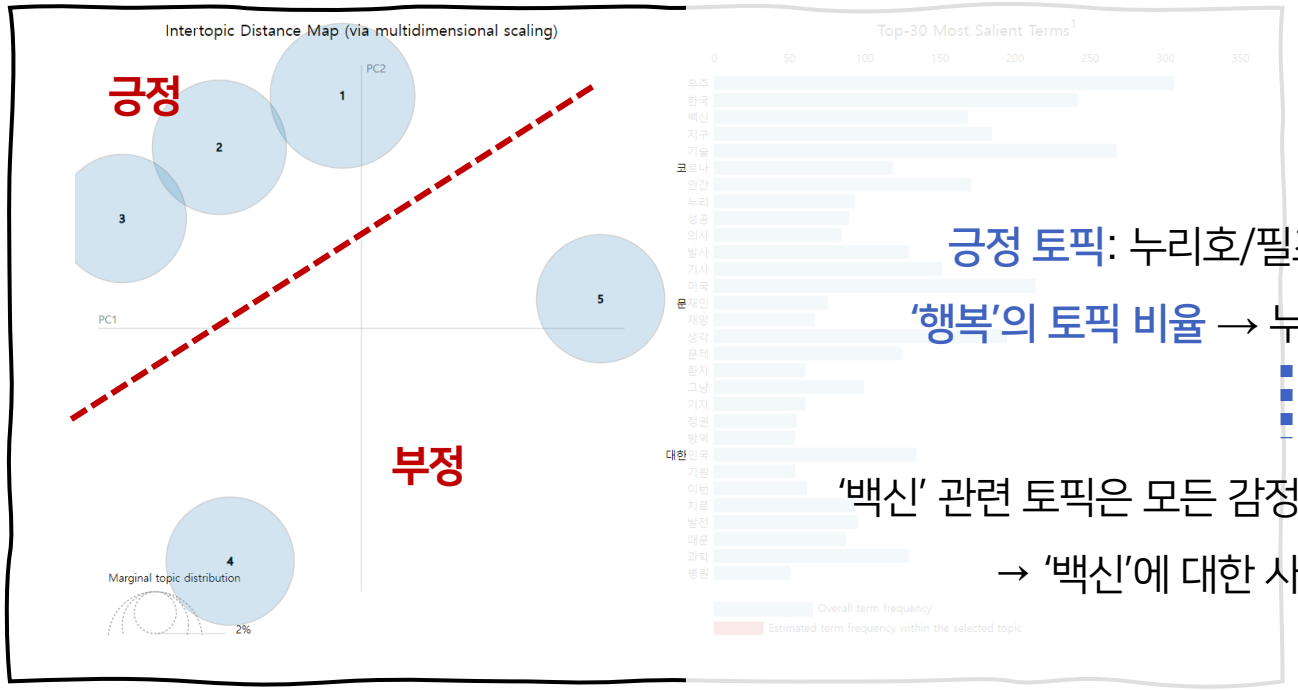



같은 방식으로 과학 분야 또한 분석

**대각선 기준**으로 크게 긍/부정으로 분류 가능

## 결과 해석

과학 전체




**긍정 토픽**: 누리호/필즈상/반도체 기술 등

**'행복'의 토픽 비율** → 누리호/필즈상이 대부분

'백신' 관련 토픽은 모든 감정마다 **독립적 토픽**으로 존재

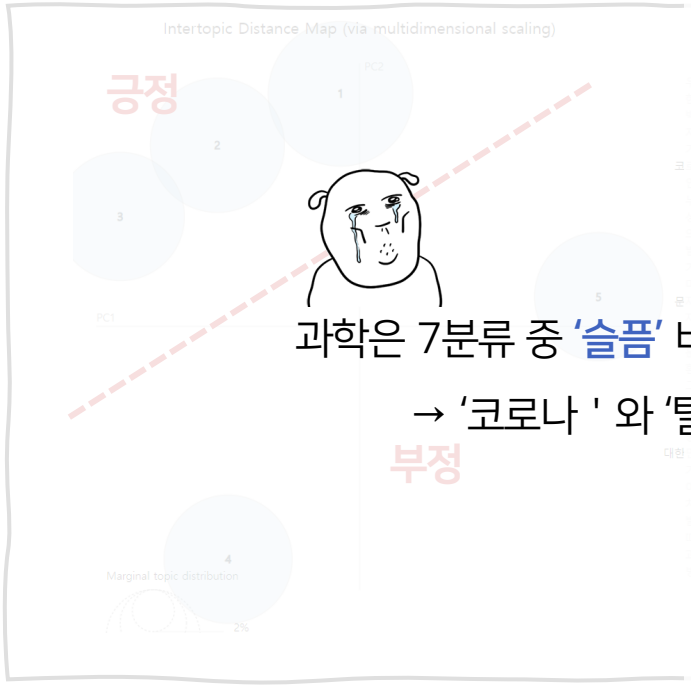
→ '백신'에 대한 사람들의 의견 분분

같은 방식으로 과학 분야 또한 분석

**대각선 기준**으로 크게 긍/부정으로 분류 가능

## 결과 해석

과학 전체






과학은 7분류 중 '슬픔' 비율이 두 번째로 높음

→ '코로나 ' 와 '탈모'가 그 이유

## 결과 분석

### 행복을 드러내는 양상

대부분 2개의 토픽을 가짐 → 축하, 응원, 감사 등 유사한 양상(키워드)을 보임



### 정치 관련 토픽

모든 카테고리에서 정치 관련 토픽이 존재 → 대부분 **부정적인(분노/혐오)** 반응.

욕설로 탐지되었던 단어들이 키워드로 선택되는 경우 多

⋮

정치/경제/세계 카테고리: 정치 분야가 대부분의 토픽을 차지, 토픽의 구분이 명확 X

정치적 입장이 극명하게 다른 사용자들이 많음 → 같은 토픽임에도 단어 선택이 다름 → 구분이 모호

## 결과 분석



### 행복을 드러내는 양상

대부분 2개의 토픽을 가짐 → 축하, 응원, 감사 등 유사한 양상(키워드)을 보임

### 정치 관련 토픽

모든 카테고리에서 정치 관련 토픽이 존재 → 대부분 부정적인(분노/혐오) 반응.

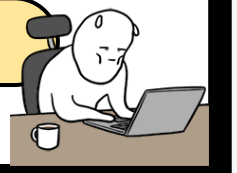
욕설로 탐지되었던 단어들이 키워드로 선택되는 경우 多

정치/경제/세계 카테고리: 정치 분야가 대부분의 토픽을 차지, 토픽의 구분이 명확 X

정치적 입장이 극명하게 다른 사용자들이 많음 → 같은 토픽임에도 단어 선택이 다름 → 구분이 모호하다고 해석

## 다음주 예고

- GAN에 대해
- DCGAN에 대해
- BERT + DCGAN
  이모티콘 생성
- Loss/그 외 성능 높이기



선형대수학

# 감사합니다