seok830621

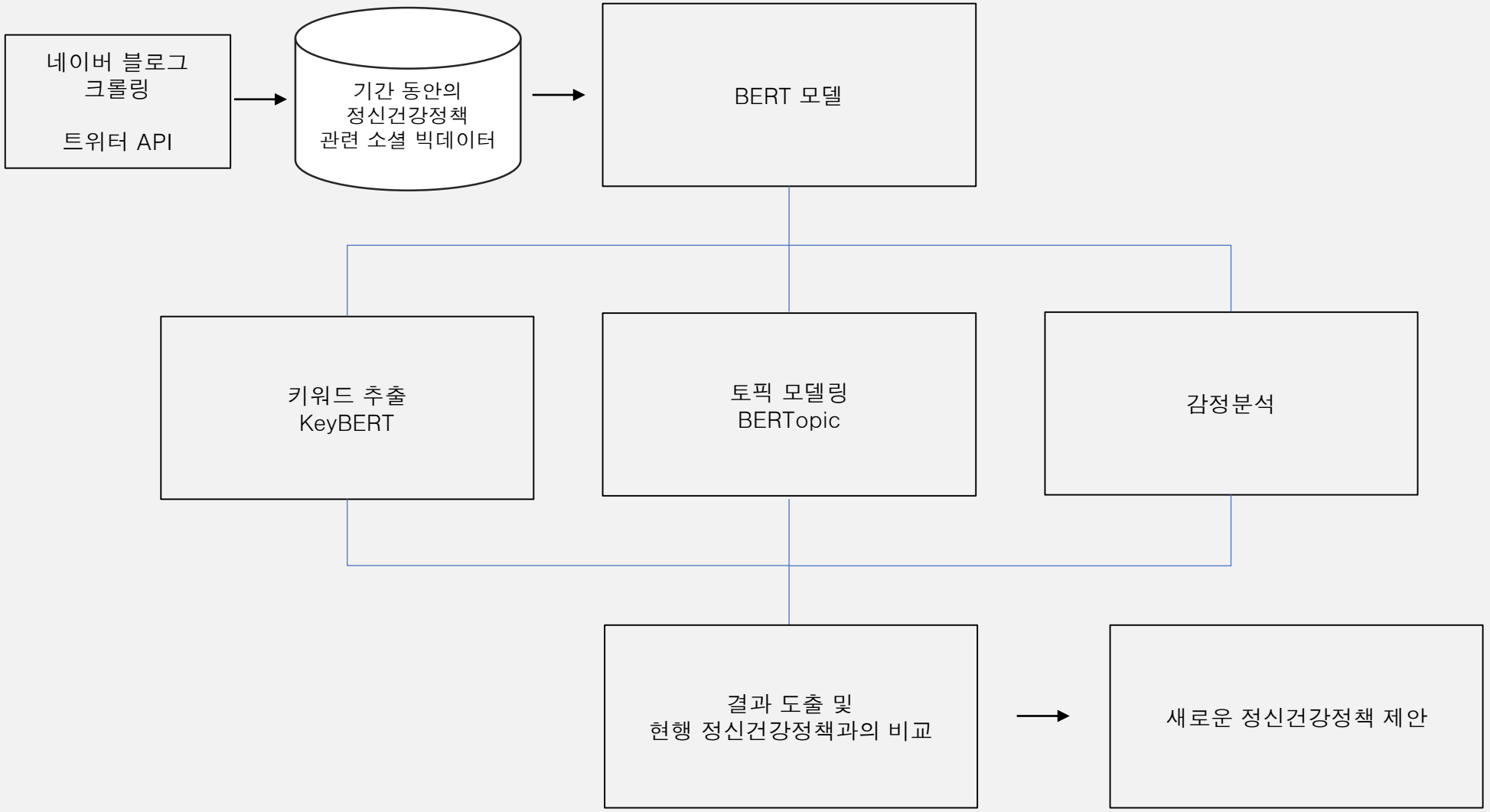
소셜 빅데이터 분석을 통한 정신건강정책 제안

머신러닝의 기초 4조

이전 발표 내용

- 연구 프로세스
- KeyBERT
- KoBERTopic

연구 프로세스



선행연구 논문

제목	발행연도	분석방법 및 사용모델	사용데이터	한계점
소셜 빅데이터 기반 보건복지 정책 미래 신호 예측	2016	EDA: 단어빈도, 문서빈도, TF-IDF 연관분석: Apriori 알고리즘 분류예측: Random Forest	- 소셜 빅데이터	- 3개월에 한정된 소셜 빅데이터 - 집단을 기준으로 분석하여 개인에게 적용할 경우 생태학적 오류가능성 - 문서내 의미와 이론적 모형 내의 의미차이 발생 가능성
기후변화 정책 수립 지원을 위한 소셜 빅데이터 분석	2019	NLP: KoNLPy, Doc2Vec(word embedding) 토픽 모델링 토픽 모델링 평가: 응집성 분석 시각화: pyLDAvis	- 소셜 빅데이터	- 데이터를 수집한 네이버 블로그의 이용자수가 적음 - 데이터를 해석 및 분석할 때 연구자의 주관 개입가능성 - 인터넷과 SNS 이용자의 의견만 수집

선행연구 논문

제목	발행연도	분석방법 및 사용모델	사용데이터	한계점
데이터 사이언스를 활용한 출산 정책 관련 여론 분석	2017	NLP: 유의어-제외어 처리, 한글 자연어처리 의미망 분석: 백본모델, 거번-뉴먼 알고리즘 키워드 분석: 보나시치 영향력 지수	- 인터뷰 데이터 - 온라인 커뮤니티 데이터	- 여성 미성년자의 삶의 질을 통합적으로 고려하지 않음 - 가설과 검증이라는 기존의 연구방법론에 적합하지 않음 - 필요한 모든 자료와 변수 수집 불가능 - 변수의 관계를 찾는 것에 한계가 있음
한국 미혼모에 대한 관점 변화와 정부정책의 방향: 1995년~2020년 소셜미디어 빅데이터 분석	2021	EDA: 단어빈도, TF-IDF, N-gram 네트워크 분석 의미연결망 분석(CONCOR분석)	- 소셜 빅데이터	- 빅데이터 활용이 높지 않았던 시기도 연구 기간에 포함하였기 때문에 당시의 사회를 대변하지 못함 - 키워드인 '미혼모', '싱글맘', '비혼모'가 시대별로 사용되었기 때문에 정확하게 비교하는 데에 한계가 있음

기후변화 정책 수립 지원을 위한 소셜 빅데이터 분석

- 연구 목적: 기후변화 정책 수요 파악
- 소셜 빅데이터: 트위터, 네이버 블로그(기후변화)
- BeautifulSoup로 게시글 텍스트 추출
- 자연어 처리: KoNLPy, Doc2Vec(word embedding)-> positive
- 기후변화: [‘기후변화‘], [‘기후변화‘, ‘심각‘], [‘기후변화’,’취약‘]
- 기후변화 정책: [‘기후변화’, ‘정책‘], [‘기후변화‘, ‘정책’, ‘우선‘]
- 토픽 모델링: LDA 기법
- 토픽 모델링 평가: 응집성 분석

기후변화 정책 수립 지원을 위한 소셜 빅데이터 분석

- 한계점
  - 데이터를 수집한 네이버 블로그의 이용자수가 적음
  - 데이터를 해석 및 분석할 때 연구자의 주관 개입가능성
  - 인터넷과 SNS 이용자의 의견만 수집

seok830621

뉴스 크롤링

In [1]:
```
#크롤링시 필요한 라이브러리 불러오기
from bs4 import BeautifulSoup
import requests
import re
import datetime
from tqdm import tqdm
import sys
```

In [2]:
```
# 페이지 url 형식에 맞게 바꾸어 주는 함수 만들기
  #입력된 수를 1, 11, 21, 31 ...만들어 주는 함수
def makePgNum(num):
    if num == 1:
        return num
    elif num == 0:
        return num+1
    else:
        return num+9*(num-1)
```

In [3]:
```
# 크롤링할 url 생성하는 함수 만들기(검색어, 크롤링 시작 페이지, 크롤링 종료 페이지)

def makeUrl(search, start_pg, end_pg):
    if start_pg == end_pg:
        start_page = makePgNum(start_pg)
        url = "https://search.naver.com/search.naver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&query="
        print("생성url: ", url)
        return url
    else:
        urls = []
        for i in range(start_pg, end_pg + 1):
            page = makePgNum(i)
            url = "https://search.naver.com/search.naver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&quer
            urls.append(url)
        print("생성url: ", urls)
        return urls
```

- BeautifulSoup을 이용한 크롤링

seok830621

뉴스 크롤링

In [4]:
```
# html에서 원하는 속성 추출하는 함수 만들기 (기사, 추출하려는 속성값)
def news_attrs_crawler(articles,attrs):
    attrs_content=[]
    for i in articles:
        attrs_content.append(i.attrs[attrs])
    return attrs_content

# ConnectionError방지
headers = {"User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) Chrome/98.0.4758.102"}
```

In [5]:
```
#html생성해서 기사크롤링하는 함수 만들기(url): 링크를 반환
def articles_crawler(url):
    #html 불러오기
    original_html = requests.get(i,headers=headers)
    html = BeautifulSoup(original_html.text, "html.parser")

    url_naver = html.select("div.group_news > ul.list_news > li div.news_area > div.news_info > div.info_group > a.info")
    url = news_attrs_crawler(url_naver,'href')
    return url
```

In [6]:
```
#####뉴스크롤링 시작#####

#검색어 입력
search = input("검색할 키워드를 입력해주세요:")
#검색 시작할 페이지 입력
page = int(input("₩n크롤링할 시작 페이지를 입력해주세요. ex)1(숫자만입력):")) # ex)1 =1페이지,2=2페이지...
print("₩n크롤링할 시작 페이지: ",page,"페이지")
#검색 종료할 페이지 입력
page2 = int(input("₩n크롤링할 종료 페이지를 입력해주세요. ex)1(숫자만입력):")) # ex)1 =1페이지,2=2페이지...
print("₩n크롤링할 종료 페이지: ",page2,"페이지")
```

```
검색할 키워드를 입력해주세요:정신건강정책

크롤링할 시작 페이지를 입력해주세요. ex)1(숫자만입력):1

크롤링할 시작 페이지:  1 페이지

크롤링할 종료 페이지를 입력해주세요. ex)1(숫자만입력):29999

크롤링할 종료 페이지:  29999 페이지
```

- 검색 키워드: 정신건강정책
- 1페이지 ~ 29,999 페이지

뉴스 크롤링

```
# naver url 생성
url = makeUrl(search,page,page2)

#뉴스 크롤러 실행
news_titles = []
news_url =[]
news_contents =[]
news_dates = []
for i in url:
    url = articles_crawler(url)
    news_url.append(url)


#제목, 링크, 내용 1차원 리스트로 꺼내는 함수 생성
def makeList(newlist, content):
    for i in content:
        for j in i:
            newlist.append(j)
    return newlist


#제목, 링크, 내용 담을 리스트 생성
news_url_1 = []

#1차원 리스트로 만들기(내용 제외)
makeList(news_url_1,news_url)

#NAVER 뉴스만 남기기
final_urls = []
for i in tqdm(range(len(news_url_1))):
    if "news.naver.com" in news_url_1[i]:
        final_urls.append(news_url_1[i])
    else:
        pass
```

seok830621

뉴스 크롤링

```
# 뉴스 내용 크롤링

for i in tqdm(final_urls):
    #각 기사 html get하기
    news = requests.get(i,headers=headers)
    news_html = BeautifulSoup(news.text,"html.parser")

    # 뉴스 제목 가져오기
    title = news_html.select_one("#ct > div.media_end_head.go_trans > div.media_end_head_title > h2")
    if title == None:
        title = news_html.select_one("#content > div.end_ct > div > h2")

    # 뉴스 본문 가져오기
    content = news_html.select("div#dic_area")
    if content == []:
        content = news_html.select("#articeBody")

    # 기사 텍스트만 가져오기
    # list합치기
    content = ''.join(str(content))

    # html태그제거 및 텍스트 다듬기
    pattern1 = '<[^>]*>'
    title = re.sub(pattern=pattern1, repl='', string=str(title))
    content = re.sub(pattern=pattern1, repl='', string=content)
    pattern2 = """[₩n₩n₩n₩n// flash 오류를 우회하기 위한 함수 추가₩nfunction _flash_removeCallback() {}"""
    content = content.replace(pattern2, '')

    news_titles.append(title)
    news_contents.append(content)

    try:
        html_date = news_html.select_one("div#ct> div.media_end_head.go_trans > div.media_end_head_info.nv_notrans > div.medi
        news_date = html_date.attrs['data-date-time']
    except AttributeError:
        news_date = news_html.select_one("#content > div.end_ct > div > div.article_info > span > em")
        news_date = re.sub(pattern=pattern1,repl='',string=str(news_date))
    # 날짜 가져오기
    news_dates.append(news_date)
```

뉴스 크롤링

```
print("검색된 기사 갯수: 총 ",(page2+1-page)*10,'개')
print("₩n[뉴스 제목]")
print(news_titles)
print("₩n[뉴스 링크]")
print(final_urls)
print("₩n[뉴스 내용]")
print(news_contents)

print('news_title: ',len(news_titles))
print('news_url: ',len(final_urls))
print('news_contents: ',len(news_contents))
print('news_dates: ',len(news_dates))

###데이터 프레임으로 만들기###
import pandas as pd

#데이터 프레임 만들기
news_df = pd.DataFrame({'date':news_dates,'title':news_titles,'link':final_urls,'content':news_contents})

#중복 행 지우기
news_df = news_df.drop_duplicates(keep='first',ignore_index=True)
print("중복 제거 후 행 개수: ",len(news_df))

#데이터 프레임 저장
now = datetime.datetime.now()
news_df.to_csv('{}_{}.csv'.format(search,now.strftime('%Y%m%d_%H시%M분%S초')),encoding='utf-8-sig',index=False)
```

```
aver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&query=정신건강정책&sort=1&start=299861', 'https://sea
rch.naver.com/search.naver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&query=정신건강정책&sort=1&start
=299871', 'https://search.naver.com/search.naver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&query=정
신건강정책&sort=1&start=299881', 'https://search.naver.com/search.naver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom202111
01to20221031&&query=정신건강정책&sort=1&start=299891', 'https://search.naver.com/search.naver?where=news&sm=tab_pge&nso=s
o%3Add%2Cp%3Afrom20211101to20221031&&query=정신건강정책&sort=1&start=299901', 'https://search.naver.com/search.naver?where
=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&query=정신건강정책&sort=1&start=299911', 'https://search.naver.
com/search.naver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&query=정신건강정책&sort=1&start=299921',
'https://search.naver.com/search.naver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&query=정신건강정책&
sort=1&start=299931', 'https://search.naver.com/search.naver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to2022103
1&&query=정신건강정책&sort=1&start=299941', 'https://search.naver.com/search.naver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3
Afrom20211101to20221031&&query=정신건강정책&sort=1&start=299951', 'https://search.naver.com/search.naver?where=news&sm=tab
_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&query=정신건강정책&sort=1&start=299961', 'https://search.naver.com/search.n
aver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&query=정신건강정책&sort=1&start=299971', 'https://sea
rch.naver.com/search.naver?where=news&sm=tab_pge&nso=so%3Add%2Cp%3Afrom20211101to20221031&&query=정신건강정책&sort=1&start
=299981']
```

```
100%|██████████████████████████████████████████████████| 390099/390099 [00:00<00:00, 2452497.92it/s]
 42%|████████████████████▍                              | 37441/90109 [1:24:45<2:11:20,  6.68it/s]
```

```
print('a')
```

수집된 데이터

```
[ ]  data
```

	date	title	link	content
0	2022-10-31 20:02:01	"내 딸, 내 아들 같은 안타까움에…집단 트라우마 노출 위험"…전문가들 "영상·…	https://n.news.naver.com/mnews/article/366/000…	[₩n일상적 공간서 벌어진 참사가 과몰입 유도트라우마 반복 노출 시 PTSD 걸릴 …
1	2022-10-31 18:21:03	부상자·유가족 치료·장례 지원 [이태원 비극 수습 속도전]	https://n.news.naver.com/mnews/article/014/000…	[₩n복지부 '사고수습본부' 설치직원 파견 일대일 매칭 관리 보건복지부가 10월 …
2	2022-10-31 17:33:03	[매경춘추] 한 해를 마무리할 때	https://n.news.naver.com/mnews/article/009/000…	[₩n₩n₩n₩n₩n 2022년 달력이 달랑 두 장 남았습니다. 지금까지 무사히 버…
3	2022-10-31 17:31:01	한국상담심리학회, 이태원 참사 수습 지원 위한 대응팀 운영	https://n.news.naver.com/mnews/article/119/000…	[₩n₩n₩n₩n₩nⓒ[데일리안 = 이현남 기자] (사)한국상담심리학회(학회장 이동…
4	2022-10-31 15:55:01	경제력과 학력 차이가 코로나로 인한 건강 불평등 악화시켜	https://n.news.naver.com/mnews/article/018/000…	[₩n서울대병원, 코로나 전후 주관적 건강 및 사회경제적요인 간 연관강도 차이 규명…
…	…	…	…	…
1307	2022-05-17 06:02:00	서울시 '생명사랑 키트' 국제디자인 어워드 디자인상	https://n.news.naver.com/mnews/article/277/000…	[₩n응급실 내원한 자살시도자 사후관리 위한 '생명사랑 키트', 서비스디자인 부문 …
1308	2022-05-17 06:01:14	서울시 '생명사랑 키트', 국제디자인 어워드 수상	https://n.news.naver.com/mnews/article/001/001…	[₩n₩n₩n₩n₩n서울시 '생명사랑 키트'[서울시 제공. 재판매 및 DB 금지] …
1309	2022-05-17 06:01:05	서울시가 만든 '생명사랑 키트', IF 어워드 디자인상 수상	https://n.news.naver.com/mnews/article/014/000…	[₩n₩n₩n₩n₩n사진=서울시[파이낸셜뉴스] 서울시는 '생명사랑 키트(사진)'가 …
1310	2022-05-17 06:00:00	'생명사랑 키트-7일간의 도전' iF어워드 디자인상 수상	https://n.news.naver.com/mnews/article/003/001…	[₩n기사내용 요약자살시도자 사후관리 위한 '생명사랑 키트' if 어워드서 서비스 …
1311	2022-05-16 17:11:03	억만장자가 삼키는 공론장…트위터, 혐오에 날개 다나	https://n.news.naver.com/mnews/article/028/000…	[₩n머스크 인수에 우려 목소리머스크는 "언론 자유 최대 확장"인수 땐 규제 풀어 …

1312 rows × 4 columns

KoBERTopic

```
[ ] class CustomTokenizer:
        def __init__(self, tagger):
            self.tagger = tagger
        def __call__(self, sent):
            sent = sent[:1000000]
            word_tokens = self.tagger.morphs(sent)
            result = [word for word in word_tokens if len(word) > 1]
            return result
```

```
[ ] custom_tokenizer = CustomTokenizer(Mecab())
```

```
[ ] vectorizer = CountVectorizer(tokenizer=custom_tokenizer, max_features=3000)
```

```
[ ] docs
```

['| 일상적 공간서 벌어진 참사가 과몰입 유도트라우마 반복 노출 시 PTSD 걸릴 수 있어 "슬픔은 당연한 반응…방아돌여야 뇌 건강" 31일 서울 용산구 이태원 압사사고 인근에 마련된 추모공간을 찾은 시민들이 희생자를 추모하고 있다. /뉴스1 지난 29일 서울 용산구 이태원 해밀턴 호텔 옆 골목에서 발생한 대규모 압사 사고로 154명이 숨지고 149명이 다친 것으로 집계됐다. 희생자들은 주말 할러윈을 맞이해 몰려 나온 10대와 20대가 주를 이뤘다.사고 당시 현장을 담은 영상과 사진은 유튜브를 비롯한 소셜미디어(SNS)를 통해 확산하고 있다. 트위터코리아는 이태원 사고 현장 사진, 영상을 공유할 시 내부 정책에 따라 제재를 받을 수 있다고 경고했다. 문제가 되는 게시물을 발견하면 신고해달라는 공지도 남겼다.이와 관련해 정신과 관련 학회들은 사고 상황을 기록한 영상·사진 공유를 중단해달라 촉구하는 내용의 성명을 발표하고 있다. 30일 대한신경정신의학회에 이어 31일에는 정신건강의학과의사회가 성명서를 발표했다.정신건강의학과의사회는 성명서에서 "(이태원) 사건 현장을 담은 영상과 사진이 트라우마로 작용해 외상 후 스트레스 장애(PTSD)를 일으킬 수 있다"며 "과도한 몰입은 자제하는 게 좋다"고 경고했다.트라우마(Trauma)는 인간이 일상생활이나 특수한 상황 속에서 겪을 수 있는 극단적이거나 충격적인 사건이다. 교통사고, 건물 붕괴 사고, 지진이나 태풍과 같은 자연재해, 암과 같은 질병 등 생명의 위협을 느끼게 하는 심각한 사건이다. 이태원 사건과 같은 폭발적 정신적 상해를 남기는 일이나, 현장을 담은 영상, 사진 모두 트라우마가 된다.전문가들은 이태원 사고 이후 수많은 영상과 사진이 확산하면서 많은 국민이 광범위한 트라우마에 노출된 상태라고 진단한다. 이는 이태원 사고가 가진 특수성 때문이다. 이많은 정신과 전문의는 "해외에서 벌어지는 전쟁이나 자연재해 등으로 아무리 많은 사상자가 나와도 우리에게 별다른 정신적 영향이 없다"며 "'자신과 상관 없는 일'이라는 생각이 무의식 중에 깔려있기 때문"이라고 설명했다.반면 이태원 사고는 많은 사람들이 평소 일상적으로 오가는 공간에서 발생했다. 정서적으로 가까운 곳에서 벌어진 참사이기 때문에 사고에 몰입하는 수준이 상대적으로 활성 높다는 것이다. 여기에 다른 사고들과 달리 시신 수습이 제대로 되지 않은 상태에서 촬영된 영상과 사진이 상당수 유포되면서 더욱 강한 트라우마로 작용할 가능성이 커졌다는 것이다.이런 트라우마에 반복 노출되는 것은 PTSD를 일으키는 주요 원인이 된다. 인간이 트라우마에 직면하면 뇌에서는 '변연계'라 불리는 부분이 평소보다 활성 활성화된다. 변연계는 원초적 공포, 동물적 욕구 등 극단적 긴장상태에서 나타나는 반응을 관장하는 부위다.트라우마를 겪은 사람은 변연계 과활성에 따른 '급성 스트레스 반응'으로 우울증, 불안장애를 보인다. 다만 이는 누구나 겪을 수 있어 오래가지 않는다. 공포 영화나 잔인한 영상을 보는 것만으로 변연계는 순간 과활성되지만 빠르게 원래 상태로 돌아오는 것과 같은 이치다.하지만 트라우마를 장기간 반복적으로 겪으면 급성 스트레스 반응이 오래 지속되면서 PTSD로 이어질 가능성이 커진다. 변연계가 과활성화된 채 원상태로 돌아오지 않으면서 뇌가 극단적 긴장을 유지하는 것이다. 그 결과 PTSD 환자에게서는 과도한 예민함, 조울증, 집중력 장애, 수면장애가 나타난다. 전문가들이 이태원 사고 영상, 사진을 반복 시청하지 말라고 권유하는 이유다.이 전문의는 "코로나 바이러스가 전염되듯 이태원 사고에 집단적으로 과몰입하면서 생기는 불안정한 정서가 점점 더 많은 국민에게 안 좋은 영향을 미칠 수 있다"며 "현장 상황을 자세히 묘사하는 모든 것들로부터 거리를 두어 한다"고 말했다.이번 사고에서 개인이 느끼는 감정을 적절히 해소할 방안을 마련해야 한다는 주장도 나온다. 홍나래 한림대성심병원 정신건강의학과교수는 "사고 관련 영상과 사진을 피하는 건 좋지만, 사고 사건을 접하여 드는 안타까움이나 슬픈 감정들을 억누르는 건 오히려 정신 건강에 해로울 수 있다"고 말했다.홍 교수는 "외부 스트레스에 적절히 반응해야 오히려 뇌 신경 전달 물질들이 균형 있게 분비될 수 있다"며 "같은 사건을 봐도 슬픔의 크기는 다를 수 있기 때문에, 다른 사람의 반응을 보며 유난을 떤다 비난하는 행위도 삼가야 한다"고 말했다. ']

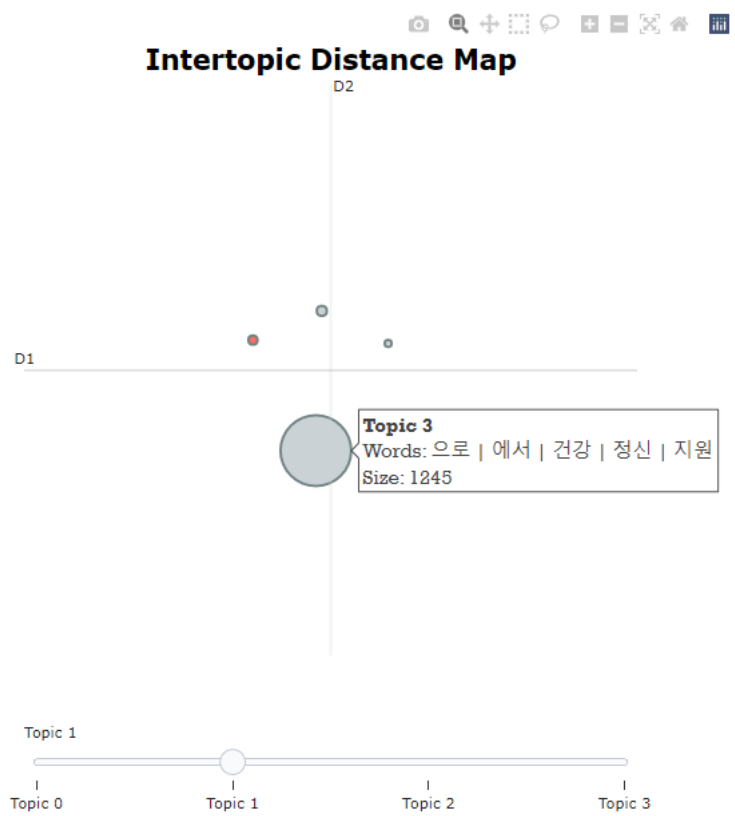
```
[ ]  model = BERTopic(embedding_model="sentence-transformers/xlm-r-100langs-bert-base-nli-stsb-mean-tokens", #
                      vectorizer_model=vectorizer,
                      nr_topics=10,
                      top_n_words=10,
                      calculate_probabilities=True)
```

- BERTopic 모델에 tokenizer는 Mecab으로 설정
- BERT모델은 다국어 SBERT로 설정

KoBERTopic

```
[ ]  topics, probs = model.fit_transform(docs)
```

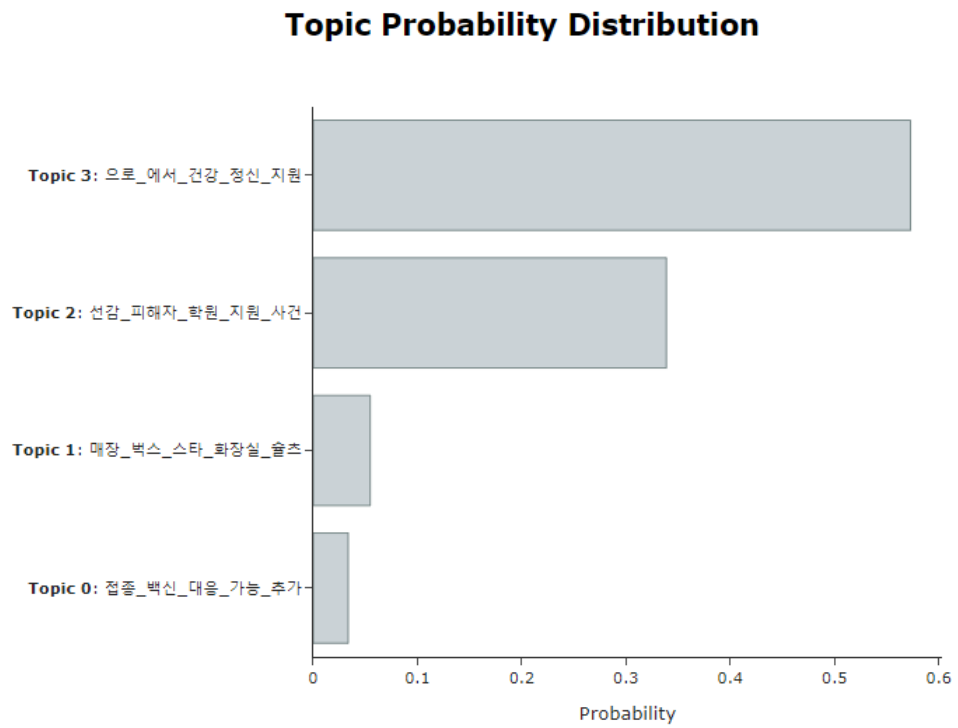
```
model.visualize_topics()
```



- Topic 도출

KoBERTopic

```
[ ]  model.visualize_distribution(probs[0])
```



- Topic Probability 도출

```
[ ]  for i in range(0, 10):
       print(i,'번째 토픽 :', model.get_topic(i))
```

```
0 번째 토픽 : [('접종', 0.3718622273507584), ('백신', 0.2102802551236464), ('대응', 0.10840297129171272), ('가능', 0.10553832937035056), ('추가', 0.09256718468685042), ('동절기', 0.09204610321814245), ('대상', 0.08918843438322485), ('시설', 0
1 번째 토픽 : [('매장', 0.2765140534169209), ('벅스', 0.20688734714590515), ('스타', 0.2013642117670126), ('화장실', 0.1889103304939464), ('슐츠', 0.11919237194550315), ('개방', 0.11280398548013203), ('ceo', 0.10980294537385334), ('직원', 0.0
2 번째 토픽 : [('선감', 0.07973068730374545), ('피해자', 0.07745742166915479), ('학원', 0.07164041475864036), ('지원', 0.05146624142822842), ('사건', 0.05024057685320422), ('에서', 0.04090603996506343), ('촉법소년', 0.040706082759120576), ('
3 번째 토픽 : [('으로', 0.04621766733753876), ('에서', 0.03997412248101475), ('건강', 0.029929218224133847), ('정신', 0.027980774793568837), ('지원', 0.027884210740037184), ('사회', 0.025208128467032855), ('정책', 0.022712594221579399), ('한다
4 번째 토픽 : False
5 번째 토픽 : False
6 번째 토픽 : False
7 번째 토픽 : False
8 번째 토픽 : False
9 번째 토픽 : False
```

이후 계획

- 블로그, 트위터 데이터 수집
- KeyBERT를 통한 키워드 추출
- 감정분석 구현
- 다른 모델과의 성능 비교
- 현 정신건강정책과의 비교 및 보완 방향