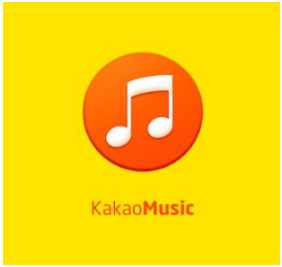
카카오톡 채팅방 감정기반 BGM 추천

Sentiment Analysis 2조
18기 분석 박영은
18기 분석 이가은
18기 분석 이소연
18기 시각화 홍주리

IDEA

과거 싸이월드의 BGM(Back Ground Music)처럼
상대방과의 채팅방에 들어갔을 때
노래가 흘러나온다면 어떨까?



내 프로필 뮤직과 어울리는 음악 듣기 100곡 >

현재 카카오톡에서
프로필 뮤직 추천 서비스를 제공 중

카카오톡 채팅방 감성에 어울리는
노래를 추천해주는 서비스가 생겼으면 좋겠다!

Table of Contents

01

카카오톡 채팅방 감성분석

BERT & KOBERT 사용

BERT(Bidirectional Encoder Representations from Transformers) 모델

- 구글이 발표한 pretrained 딥러닝 언어 모델 (2018년 말에 NLP 분야 11개 분야에서 모두 SOTA를 달성한 모델)
- 기존의 NLP 모델은 (글을 읽는 순서대로)왼쪽에서 오른쪽으로 갔지만,

  BERT는 이 순서를 양방향으로(bidirectional) 본다는 특징
- 영어 및 103개 언어에 대한 pretrained 언어 모델 제공

KoBERT 모델

- BERT 모델 구조에 한국어 위키 및 뉴스를 추가 학습 시킨 모델

TRAIN DATA

'Ai허브'에서 제공한 단발성 & 연속성 대화 데이터 셋을 활용하여 전처리 및 통합

'행복, 분노, 슬픔, 중립' label 이용하여 다중분류 모델 학습

단발성 대화

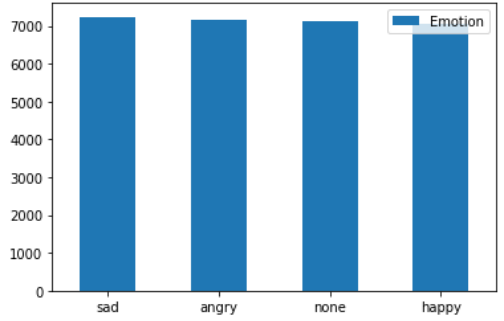
Sentence	Emotion
남자친구가 떠날까봐요	공포
내가불안해서꾸는걸까..	공포
뭐지 뜬금없이ㅋㅋㅋㅋ	놀람
워메 이게 뭔일이랴	놀람
엄마의 한계는 어디인가?	놀람
우주항공 창조혁신센타 만들어야 할듯	중립
그냥 눈도 온다고 그래~	중립
대댓글에 공감표시 허용 굿임다	행복
오늘편 레전드"~~~~~~~	행복
비와이 진짜 우와진짜 최고최고최고	행복
왜저래. 영화인줄아나?	혐오
퇴학시키고 구속시켜라	혐오

+

연속성 대화

dialog #	발화	감정
5	아 진짜 사무실에서 피지 말라니깐! 간접흡연이 얼마나 안좋은데!	분노
	그럼 직접흡연하는 난 얼마나 안좋겠니? 안그래? 보면 꼭... 지 생각만 하고.	혐오
	손님 왔어요.	중립
	손님? 누구?	중립
	클라요. 팀장님 친구래요.	중립
	내 친구? 친구 누구?	중립
	그걸 내가 어떻게 알아요!	분노
5	그래서... 무슨 일 해?	중립
	그냥 방송일 조금.	중립
	방송? 방송 뭐?	중립
	케이블 같은데서 아침에 배도 타고 산도 오르고, 있어 그런 거.	중립
	사는 덴 어디야?	중립
	개포동.	중립
	남편은 뭐하는데?	중립
	등사무소에서 인구조사 나왔니? 지금.	분노
5	언제부터 시작할 수 있어?	중립
	나? _ 나보고 하라고?	놀람
	그럼 내가 널 왜 만나러 왔겠어?	혐오
	근데 왜... 나한테...?	놀람
	내가 몇 명 만나봤는데, 보니까 죄다 사기꾼 같고 영 못 믿겠어. 아무래도 이런 일은 아는 사람한테 말기는 게... 만 사람들도 다 그러더라.	혐오

약 28000개의 Train Data 생성



BERT & KOBERT 학습

BERT / KoBERT 학습 및 평가

- Train data : Test data = 0.75 : 0.25
- BertTokenizer 로 토큰화

```
print(sentences[0])  #토크나이징 전
print(tokenized_texts[0]) #토크나이징 후
```

```
[CLS] 삭제할게요 [SEP]
['[CLS]', '삭', '##제', '##할', '##게', '##요', '[SEP]']
```

```
import pandas as pd
from transformers import BertTokenizer

tokenizer = BertTokenizer.from_pretrained("bert-base-multilingual-cased", do_lower_case=False)
result = tokenizer.tokenize('안녕하세요')
print(result)
```

모델링 결과

모델	Validation Accuracy
BERT	약 0.66
KoBERT	약 0.71

✔ Accuracy가 더 높은 KoBERT 모델 사용

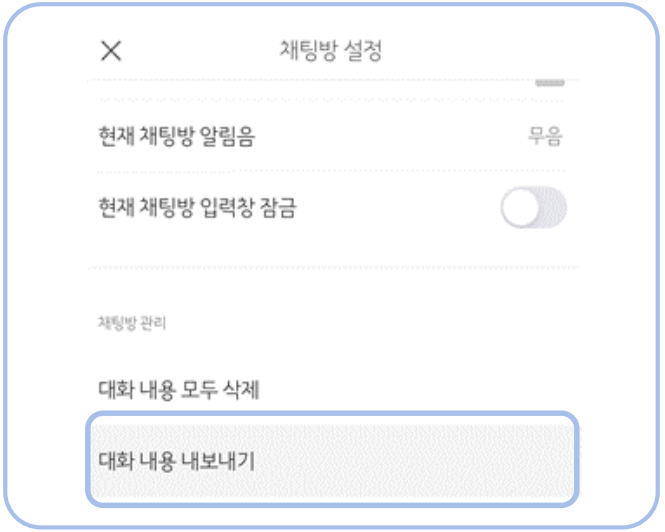
카카오톡 채팅방 데이터 불러오기

Test DATA

- 감성분석 2조 채팅방

DATA 수집기간

- ~2022년 6월 28일



모바일 카카오톡에서 '대화 내용 내보내기'
PC 카카오톡 내보내기는 적용x

카카오톡 채팅방 데이터 불러오기

```
def katalk_msg_parse(file_path):
    my_katalk_data = list()
    katalk_msg_pattern = "[0-9]{4}[년.] [0-9]{1,2}[월.] [0-9]{1,2}[일.] 오#S [0-9]{1,2}:[0-9]{1,2},.*:"
    date_info = "[0-9]{4}년 [0-9]{1,2}월 [0-9]{1,2}일 #S요일"
    in_out_info = "[0-9]{4}[년.] [0-9]{1,2}[월.] [0-9]{1,2}[일.] 오#S [0-9]{1,2}:[0-9]{1,2}:.*"

    for line in open(file_path,'rt', encoding='UTF8'):
        if re.match(date_info, line) or re.match(in_out_info, line):
            continue
        elif line == '#n':
            continue
        elif re.match(katalk_msg_pattern, line):
            line = line.split(",")
            date_time = line[0]
            user_text = line[1].split(" : ", maxsplit=1)
            user_name = user_text[0].strip()
            text = user_text[1].strip()
            my_katalk_data.append({'date_time': date_time,
                                   'user_name': user_name,
                                   'text': text
                                   })

        else:
            if len(my_katalk_data) > 0:
                my_katalk_data[-1]['text'] += "#n"+line.strip()

    my_katalk_df = pd.DataFrame(my_katalk_data)

    return my_katalk_df
```

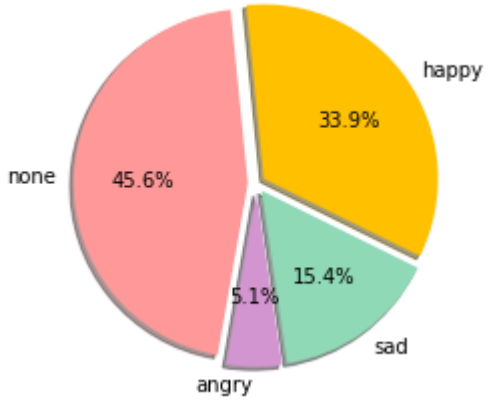
722개의 카카오톡 대화 데이터셋 생성

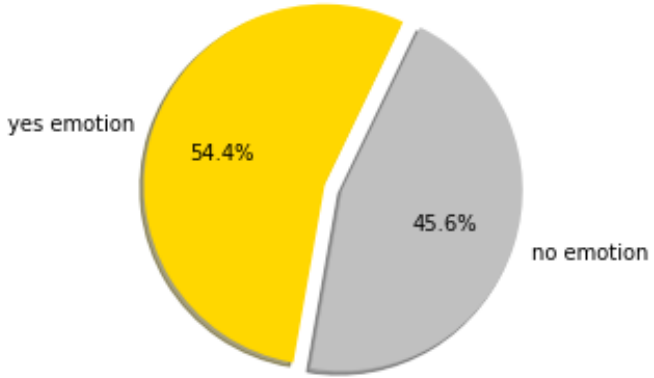
	date_time	user_name	text
0	2022. 5. 4. 오후 3:30	박영은	안녕하세요 보아즈 18기 분석 박영은입니다! 저희 다음주 세션부터 논문 발제해야해서...
1	2022. 5. 4. 오후 3:32	소연	안녕하세요😊
2	2022. 5. 4. 오후 3:34	.	안녕하세요!
3	2022. 5. 4. 오후 3:37	이가은	안녕하세요!
4	2022. 5. 4. 오후 3:43	박영은	감성분석팀이 2개조라서 논문을 지수님이 올려주신 것 두개를 나눠서 발제해야할 것 같...
...	...	...	...
718	2022. 6. 28. 오전 8:53	소연	구경해야징
719	2022. 6. 28. 오전 8:54	소연	나 아까 lstm 전까지 대본 쓰던거 일단 보낼겡
720	2022. 6. 28. 오전 8:54	소연	파일: 감성분석.pptx
721	2022. 6. 28. 오전 9:00	주리	응 여기서부터 내가할게!!!! 수고해썹 😀
722	2022. 6. 28. 오전 11:29	이가은	이모티콘

카카오톡 채팅방 감성분석

- 학습시킨 KoBert를 이용하여 카카오톡 채팅방의 모든 대화의 감정을 분류
- 행복+슬픔+분노 → 감정적 대화 / 중립 → 감정x : 대화가 얼마나 감정적인지 파악
- 감정적 대화의 비율(행복 : 슬픔 : 분노)을 노래 추천에 이용

'감성분석 2조' 채팅방 감성분석 결과





카카오톡 채팅방 감성분석

문장찾기

```
#궁금한 문장 찾기
final_data=pd.concat([df,emotion_df],axis=1)

sentence = 1
while sentence != 'End' :
    sentence = input("문장 찾기 : ")

    if sentence == 'End' :
        break

    print(final_data.text[final_data.감정==sentence])
    print("₩n")
```

```
문장 찾기 : 행복
```

행복

```
문장 찾기 : 행복
0      안녕하세요 보아즈 18기 분석 박영은입니다! 저희 다음주 세션부터 논문 발제해야해서...
1                                                안녕하세요😊
2                                                안녕하세요!
3                                                안녕하세요!
8                                           우왕 감사합니다!!
                                ...
335                                 ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
336                                  감성리뷰ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ
337                                       그치만 자연스러웠어
338                                    앗 내가 보낼게~~~!!!!
339                                       헛 고마워~~~~
Name: text, Length: 119, dtype: object
```

슬픔

```
문장 찾기 : 슬픔
6                                  저도 못받았어요! 저희가 먼저 연락해봐야할것같아요 ㅎㅎ
22                                              헉 그러네요...
37                                          삭제된 메시지입니다.
59                                                 넵👍👍
66                                                  넵
72                                          삭제된 메시지입니다.
78                                  앗 제가 화요일에는 23 이후 밖에 안돼서요ㅠㅠ
79                          제가 좀 헷갈리게 적었네요... 모두 오후 기준입니다..!!ㅠㅠ
82                                                  넵
107                                          나누기 어렵네여...
109                                  요번 논문은 진짜 길어서 나누기가 힘드네요 ㅠㅠ
131   헉!! ㅜㅠ 아쉽네요ㅠㅠ 활동 많이 하면서 친해지고 싶었눈데 😠😠₩n해외가서 좋은 ...
132   으앗 이제 친해지는줄 알았는데 많이 아쉽네염ㅜㅜㅜ 담에 좋은 기회있으면 또 뵐 수 ...
135                                          아니며 들어가요 ㅜㅜㅜㅜ
137                                              시간이 다돼서
139                                          핸드폰이에요 잠시 ㅜㅜ
```

02

노래가사 감성분석

노래가사 데이터 구축

① 노래 가사 크롤링

멜론에서 2010년부터 2021년까지
top 100 노래 가사, 가수, 제목, 년도 크롤링

	lyrics	artist	song_name	Year
0	You don't know me	미쓰에이	Bad Girl Good Girl	2010
1	You don't know me	미쓰에이	Bad Girl Good Girl	2010
2	You don't know me	미쓰에이	Bad Girl Good Girl	2010
3	You don't know me	미쓰에이	Bad Girl Good Girl	2010
4	So shut off boy	미쓰에이	Bad Girl Good Girl	2010
...	...	...	...	...
90	끝이 아니길	릴보이 (IIIBOI)	CREDIT (Feat. 염따, 기리보이, Zion.T)	2021
91	I wish I make it till the last day	릴보이 (IIIBOI)	CREDIT (Feat. 염따, 기리보이, Zion.T)	2021
92	I wish I make it till the last day	릴보이 (IIIBOI)	CREDIT (Feat. 염따, 기리보이, Zion.T)	2021
93	I wish I make it till the last day	릴보이 (IIIBOI)	CREDIT (Feat. 염따, 기리보이, Zion.T)	2021
94	I wish I make it till the last	릴보이 (IIIBOI)	CREDIT (Feat. 염따, 기리보이, Zion.T)	2021

67275 rows × 4 columns

✓ 총 1035 곡

✓ 나머지 65곡은 연령제한으로 가져오지 못함

② 영어가사 번역 진행

- Google Translate Rest API 이용

lyrics		trans_lyrics
You don't know me	→	넌 날 몰라

- 한글과 영어가 섞인 가사의 경우 번역이 불가

lyrics		trans_lyrics
겉으론 Bad girl	✕	겉으론 Bad girl

➢ 해결 방법

- 영어가사부문만 삭제 (ex. '겉으론 Bad girl' →'겉으론')
  → 문맥의 흐름이 이상해질 것이라 판단
- 영어,한글 섞인 가사 전체를 삭제 ✔

노래가사 감성사전 Word2Vec

- 노래가사 데이터셋에는 감정이 라벨링 되어있지 않음
- 각각의 노래가사에 '행복', '분노', '슬픔' 감정을 라벨링을 하기 위하여 노래가사 감성 사전 구축
- 노래 가사의 경우 비유적 혹은 은유적 표현을 많이 사용
  → 단어간 거리 유사도를 기반으로 사전을 구축하기 위하여 Word2Vec 모델을 이용
- 노래 가사 전처리(숫자와 특수문자 제거, 명사 추출, 한 글자 제거, 불용어 제거) 후 word2vec으로 감정 키워드 추출

노래가사 감성사전 Word2Vec

```
model.wv.similar_by_word('행복', topn=50)
```

```
[('기분', 0.998372495174408),
 ('친구', 0.9983059167861938),
 ('내게', 0.998293399810791),
 ('이젠', 0.9982684254646301),
 ('우리', 0.9981916546821594),
 ('하나', 0.9981793761253357),
 ('눈물', 0.998175621032714 8),
```

```
model.wv.similar_by_word('슬픔', topn=50)
```

```
[('내게', 0.997450590133667),
 ('때문', 0.9973227977752686),
 ('어디', 0.9972277879714966),
 ('이별', 0.997195124626159 7),
 ('머리', 0.997180223464965 8),
 ('그대', 0.997165977954864 5),
 ('아침', 0.997140645980835),
```

```
model.wv.similar_by_word('짜증', topn=50)
```

```
[('선택', 0.9643637537956238),
 ('습관', 0.9640271663665771),
 ('저기', 0.9634497165679932),
 ('걸음', 0.9634368419647217),
 ('플라이', 0.9633605480194092),
 ('남자', 0.9631988406181335),
 ('파티', 0.9631829261779785),
```

- 유사도가 0.75 이상인 단어 중 '내게'와 같이 중복이 되는 단어들은 제거
- 분노라는 단어가 가사에 존재하지 않아 짜증으로 대체

Problem

중복되는 단어가 많아 사전 구축에 어려움을 겪음

Solution

Word2Vec을 통해 나온 유사도 높은 단어들을 기반으로 수작업으로 감성 사전 재구성

노래가사 감성사전 Word2Vec

생성된 노래가사 감성사전

```
sad = ['그리움', '외로움', '욕심', '삶', '몫', '고통', '흔적', '인생', '심장', '방황', '그립', '혼자', '그리워','후회','부족','떠나지마','눈물','소주','이별']
angry = ['한심', '살만', '덤덤', '허무', '처량', '야속', '모욕', '불행', '비겁','다툼','각자','상처','화']
happy = ['아름답', '사랑', '영원', '고백', '설레', '매력', '감싸', '아름다운', '영원히','행복','아름다워','놀자','재미','기쁨']
```

- 위의 단어가 가사에 포함되는 경우의 빈도수 계산
- 각 가사에 대해 **빈도수가 가장 높은 감정**을 해당 가사의 감정으로 판단, 라벨링
- 2414개의 train data 구축
- 라벨링 할 감정이 없었던 가사들에 대해서는 train data를 학습시켜 라벨 예측

Train data

	trans_lyrics	감정
0	사랑하기만해도 시간 없는데	1
1	정말 행복할텐데	1
2	정말 화낼지 몰라	2
3	사랑하다 말거라면 안 할 이야기	1
4	화가 나도 소리 쳐도	2
...	...	...
2409	사랑해도 말 못 했던 나	1
2410	고마워요 행복합니다	1
2411	왜 이리 눈물이 나요	0
2412	음이 나간 내 인생은	0
2413	아쉬워 눈물이 흘러	0

2414 rows × 2 columns

0=슬픔
1=행복
2=분노

LSTM & BiLSTM 노래가사 감성분석

LSTM 모델

- LSTM 모델은 기존의 RNN이 출력과 먼 위치에 있는 정보를 기억할 수 없다는 단점을 보완하여 장/단기 기억을 가능하게 설계한 신경망의 구조
- 주로 시계열 처리나, 자연어 처리에 사용

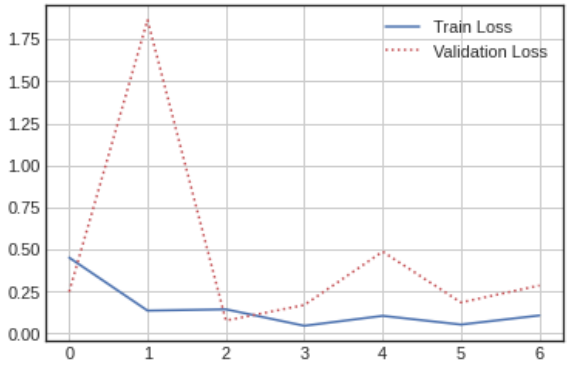
BiLSTM 모델 (양방향 LSTM)

- 시계열 또는 시퀀스 데이터의 시간 스텝 간의 양방향 장기 종속성을 학습
- 이러한 종속성은 신경망이 각 시간 스텝에서 전체 시계열로부터 학습하도록 하려는 경우에 유용

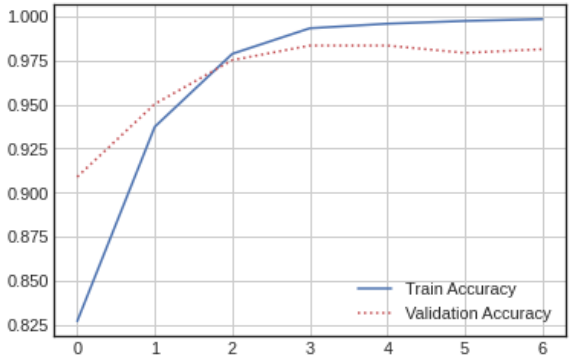
토큰화, 인코딩, 패딩의 전처리과정을 거친 Train data를 LSTM과 BiLSTM을 이용하여 가사의 감정을 분류할 수 있는 분류기 생성

LSTM & BiLSTM 노래가사 감성분석

LSTM Validation Accuracy : 0.9814



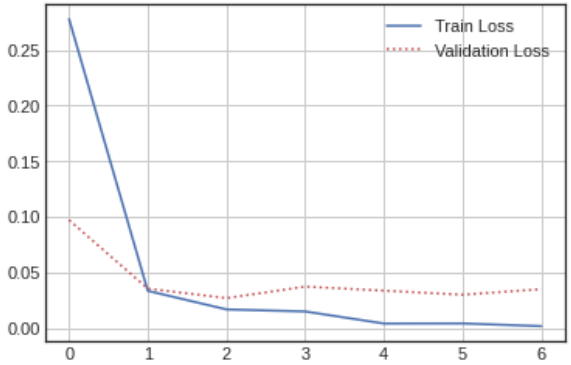
Loss



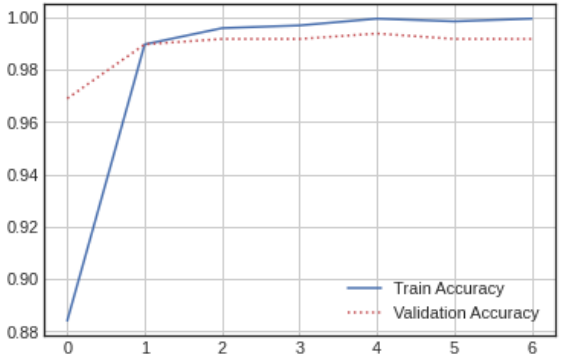
Accuracy

LSTM & BiLSTM 노래가사 감성분석

BiLSTM Validation Accuracy : 0.9917



Loss



Accuracy

BiLSTM 노래가사 감성분석

- BiLSTM으로 모든 노래 가사의 감정을 예측
- 각각의 노래 가사들이 차지하는 감정 비율을 matrix로 표현

sad	happy	angry	song_name	artist
0.516129	0.451613	0.032258	Bad Girl Good Girl	미쓰에이
0.416667	0.458333	0.125000	잔소리 (With 2AM 슬응)	아이유
0.593750	0.375000	0.031250	죽을 만큼 아파서 (Feat. 멜로우)	MC몽
0.666667	0.333333	0.000000	못해 (Feat. 美)	포맨 (4MEN)
0.750000	0.250000	0.000000	죽어도 못 보내	2am

03

채팅방 BGM추천

채팅방 BGM 추천

감성분석 2조 채팅방에 어울리는 노래는?

행복	슬픔	분노
0.6234	0.2825	0.0941

* 앞서 구한 감정 비율 중 중립을 제외하고 비율 재구성

Matrix형태의 감정 비율 중

코사인유사도를 이용해

가장 유사한 감정 비율의 노래 추천

채팅방 BGM 추천

감성분석 2조 채팅방에 어울리는 노래는?

song_name	artist
가수가 된 이유	신용재 (2F)
우리 사랑했잖아	다비치, 티아라
Love	CNBLUE (씨엔블루)
내가 노래를 못해도	SE7EN
돌덩이	하현우 (국카스텐)
I Love You	AKMU (악뮤)
Apple (Feat. 박재범)	가인
고백	양다일
너를 위해	첸 (CHEN), 백현 (BAEKHYUN), 시우민 (XIUMIN)
꼭 한번 만나고 싶다	바이브

한계

✔ 대화 데이터 감성 분석

- 카카오톡 메시지에 있는 ! (느낌표)를 주로 분노라고 구분지음
- KoBERT 모델을 fine-tuning하기 위해 필요한 데이터셋도 양이 적어, 더 많은 일상 대화 데이터를 적용시킨다면 대화 감정 분류기의 성능이 향상될 것이라고 기대됨

✔ 노래 가사 데이터 감성 분석

- 노래 가사 데이터의 양이 적어, 적절한 감정을 추출하는데 한계가 있음
- 중첩된 감정을 가진 노래 가사가 있기에, 확실한 감정 구분 짓는 것이 어려웠음

Thank You