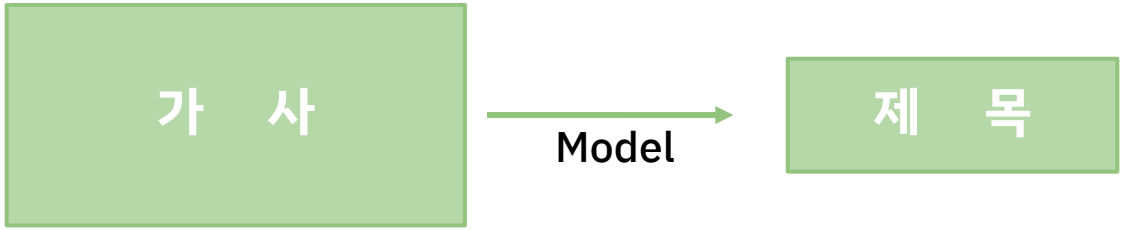
딥러닝을 이용한
한글 노래 가사 기반 제목 추천

김용재 국승용 박혜정 한민재 황성연

목 차

요 약



사용 모델

- 모델
  - LSTM + Attention
  - Transformer
  - KoBART
  - KoBERT
  - KoGPT2

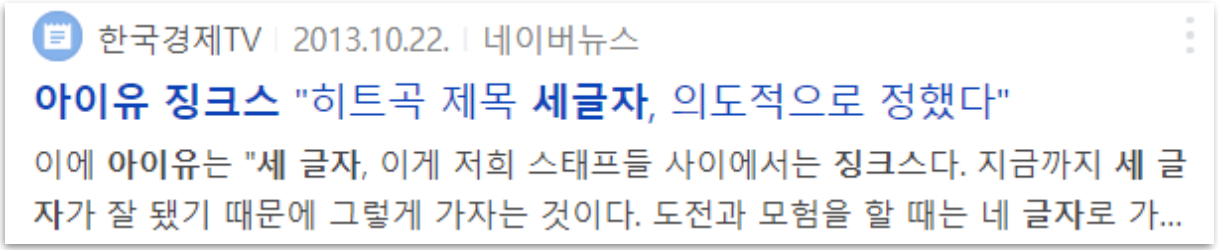
데이터 핸들링

- 전처리
  - 영어 완전 제외
  - 토큰 길이조절
- tokenizer
  - Mecab
  - OKT

2

주제 선정 배경

→ **노래 제목이 가수의 심리상태에 영향**



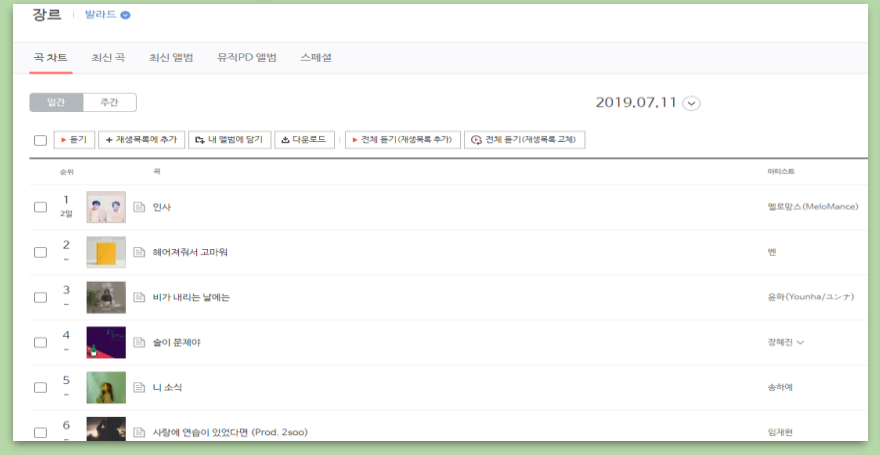
→ **노래 제목관련 흥행 징크스 有**

노래 제목이 가수의 심리상태와 흥행에 영향을 미친다는 속설이 있음
→ 가수는 제목 선정에 신경을 많이 쓰며, 그만큼 어려움이 있음

2

데이터셋

· Crawling Site



· DataFrame Set

	title	lylic
0	천만분의 1의 확률의 너 (Gravitation)	₩nEvery seconds₩r₩nEvery minutes₩r₩n너를 만나기 위해₩...
1	술이 뭐길래	₩n오늘도 이렇게 어제처럼 잔을 꺼내₩r₩n나 혼자서 따른 잔에 내 입술을 기대₩r...
2	그때 그 순간 그대로 (그그그)	₩n잘 지냈지? 조금은 어색해₩r₩n요즘 좋아 보여 인사 나누며₩r₩n사실 궁금한 ...
3	나의 X에게	₩n우리 다시 만날래₩r₩n예쁘게 빛나던₩r₩n모든 추억들이₩r₩n너무 그리워₩r₩...
4	일과 이분의 일	₩n멀리서 널 보았을 때₩r₩n다른 길로 갈까 생각했는데₩r₩n변한 듯한 널 보고 ...
...	...	...
41881	뚜벅뚜벅 (from "벨라의 꿈", 엉금이 테마)	₩n오늘도 뚜벅뚜벅₩r₩n발길은 너무나 막연해₩r₩n우리 이렇게 멀리 왔어도₩r₩n...
41882	해피 크리스마스	₩n₩r₩n₩r₩n혼자 있는 이 세상 빛이 없다 믿나요₩r₩n기댈 곳 없어 두 뺨 ...
41883	Stegosaurus	₩nOne, two, three, four, five! ₩nIt's a pentag...
41884	붕붕붕 스쿨버스 안전송	₩n헤헷! 오늘도 안전하게! 출발!₩r₩n₩r₩n스쿨비가 달려요. 붕붕붕 붕붕붕!₩...
41885	쓱싹쓱싹 클리니 청소송	₩n클리니 출동! ₩r₩n₩r₩n오케이! 준비, 오케이! 오케이! 준비, 오케이! ...

◆ 곡 500,000곡

- 멜론 : TOP100 (한국) → 10,000 곡
- 음악 블로그 : 한국 60,000 곡, 외국 40,000 곡 → 100,000 곡
- 벅스뮤직 : 한국 40,000 곡 → 40,000 곡
- 벅스뮤직 : 한국 350,000 곡 → 350,000 곡 (추가)

모델 적용 사용 모델

구 분	모델 핸들링	사전 학습 데이터
LSTM Attention	3 Layers	-
	10 Layers	-
Transformer	3 Layers	-
	10 Layers	-
KoBART	Pretrained	cosmoquester/bart-ko-small
KoBERT	Pretrained	digit82/kobart-summarization
KoGPT2	Pretrained	skt/kogpt2-base-v2

· LSTM Attention 및 Transformer : 사전 학습 X, 크롤링 데이터로 직접 학습
· KoBART, KoBERT, KoGPT2 : pretrained model 이용, 일부 학습

모델 적용 사용 모델 (LSTM Attention, Transformer)

리쌍 - 광대

사회를 살아가는 모든 사람들이 억지로 웃음을 사고 파는 '현대판' 광대라는 의미를 담고 있다.

- **LSTM Attention**

  내가 자리를 비운

- **Transformer**

  나비의 탕다

- **LSTM Attention, Transformer는 번역 결과가 기대에 미치지 못함**
- **① 입력 토큰과 출력 토큰 수의 큰 차이**
- **② 획득한 데이터로만 학습 (But, BERT · GPT → 많은 데이터로 pretrained)**

모델 적용 사용 모델 (KoBART, KoBERT, KoGPT2)

리쌍 - 광대

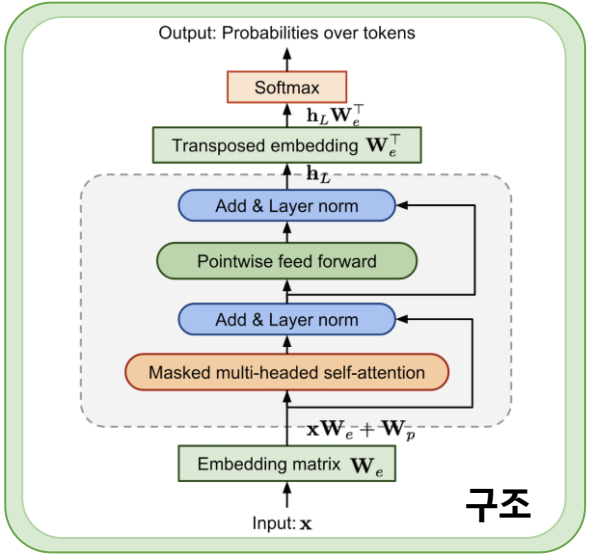
사회를 살아가는 모든 사람들이 억지로 웃음을 사고 파는 '현대판' 광대라는 의미를 담고 있다.

PICK

- **KoBART**
  - 나에게 말씀 되고
- **KoBERT**
  - 오 금강 내
- **KoGPT2**
  - 어른의 얼굴

- Transformer(KoBART), BERT(KoBERT)의 모델의 구조이해 부족
- 조건에 맞는 Output layer 추가 적용 실패
- 노래 제목 추천에 적합한 추상적 문자생성 모델 선택 : KoGPT2

모델 적용 모델 PICK : ✓ KoGPT2



학습 시킨 데이터					
</s>	<lyric>	실제 가사	<title>	실제 제목	</s>

학습된 모델로 제목 추천					
</s>	<lyric>	제목 추천 가사	<title>	추천한 제목	</s>

※ Pretrained KoGPT2 Model + 데이터 학습 Fit

- 학습 시킨 데이터는 실제 가사 및 실제 제목을 입력
- 학습된 모델로 제목 추천은 제목 추천 가사를 통해 추천한 제목을 출력

모델 적용 중간 점검 및 피드백 (KoGPT2)

예측 결과

또 하루 멀어져 간다 내뿜은 담배 연기처럼
작기만 한 내 기억 속에 무얼 채워 살고 있는지
점점 더 멀어져 간다 머물러 있는 청춘인 줄 알았는데
비어 가는 내 가슴속엔 더 아무것도 찾을 수 없네
계절은 다시 돌아오지 만 떠나간 내 사랑은 어디에
내가 떠나보낸 것도 아닌데 내가 떠나온 것도 아닌데
조금씩 잊혀져 간다 머물러 있는 사랑인 줄 알았는데
또 하루 멀어져 간다 매일 이별하며 살고 있구나

원제 : 서른즈음에
추천 제목 : 청춘은 사랑은

그대여 그대여 그대여 그대여
오늘은 우리 같이 걸어요 이 거리를 밤에 들려오는 자장
노래 어떤가요 몰랐던 그대와 단 둘이 손 잡고
알 수 없는 이 떨림과 둘이 걸어요 봄바람 휘날리며 흩날리는 벚꽃 잎이 울려 퍼질 이 거리를 둘이 걸어요 봄바람
휘날리며 흩날리는 벚꽃 잎이 울려 퍼질
이 거리를 둘이 걸어요 그대여 우리 이제 손 잡아요
이 거리에 마침 들려오는 사랑 노래 어떤가요 사랑하는 그대와 단둘이 손잡고 알 수 없는 이 거리를 둘이 걸어요

원제 : 벚꽃엔딩
추천 제목 : 봄 그리고 벚꽃 사랑

동해물과 백두산이 마르고 닳도록
하느님이 보우하사 우리나라 만세
무궁화 삼천리 화려강산
대한사람 대한으로 길이 보전하세
남산위에 저 소나무 철갑을 두른듯
바람서리 불변함은 우리 기상일세
무궁화 삼천리 화려강산
대한사람 대한으로 길이 보전하세

원제 : 애국가
추천 제목 : Milespected Switching

노랑 : 연관성 有
파랑 : 연관성 無

◆ 국내 가요는 사랑에 대한 주제 多 → 편향 문제 발생
한국어만 입력한 가사에 영어 제목이 추천되는 문제 식별

▶ 위 문제를 해결하기 위해 추가적인 핸들링 필요 판단
① 데이터 추가
② 데이터 전처리

모델 적용 데이터 추가

· 150,000 곡

- 멜론 10,000 곡
- 음악 블로그 100,000 곡
- 벅스뮤직 40,000 곡

35만곡

· 500,000 곡

- 멜론 10,000 곡
- 음악 블로그 100,000 곡
- 벅스뮤직 40,000 곡
- 벅스뮤직 350,000 곡

▶ 데이터 추가 시 기대 효과

① 예측 성능 향상
② Overfit 일부 해소
③ 데이터 다양화 (편향 일부 해소, ex. 사랑, 나)

모델 적용 데이터 전처리

구 분	과 정	비 고
#1	동명의 제목과 같은 가사들을 모두 제거	-
#2	가사의 끝부분에 필요없는 문장 제거	Bugs 님이 등록해 주신 가사입니다. \r\n\t\t\t\t\t\t\n가사 오류 제보\n
#3	제목과 가사를 한글만 남기고 지우고, 추가로 다수의 공백을 하나로 치환	정규표현식 이용
#4	형태소 분석기 mecab 사용	['천만', '분', '의', '의', '확률', '의', '너']
#5	전처리 과정에서 영어 · 숫자 제거로 조사 반복 발생 → 부자연스러운 조사 제거	['천만', '분', '의', '확률', '의', '너']
#6	제목과 가사의 토큰 개수가 1개 이하인것 제거	['흔적'] ['마침표'] ['후회']
#7	가사의 토큰수가 지나치게 적거나 많은 노래 제거	토큰 40개 이하 제거 토큰 800개 이상 제거

▶ 데이터 전처리 시 기대 효과

① **학습 데이터의 이상치 제거** → **보다 개선된 제목 추천**

② **데이터 크기를 줄임** → **한정된 컴퓨터 자원으로 효율적인 모델 학습**

결 과 KoGPT2 Model

예측 결과

또 하루 멀어져 간다 내뿜은 담배 연기처럼
작기만 한 내 기억 속에 무얼 채워 살고 있는지
점점 더 멀어져 간다 머물러 있는 청춘인 줄 알았는데
비어 가는 내 가슴속엔 더 아무것도 찾을 수 없네
계절은 다시 돌아오지 만 떠나간 내 사랑은 어디에
내가 떠나보낸 것도 아닌데 내가 떠나온 것도 아닌데
조금씩 잊혀져 간다 머물러 있는 사랑인 줄 알았는데
또 하루 멀어져 간다 매일 이별하며 살고 있구나

원제 : 서른즈음에
추천 제목 : 나의 청춘 멀어져 간다

그대여 그대여 그대여 그대여
오늘은 우리 같이 걸어요 이 거리를 밤에 들려오는 자장
노래 어떤가요 몰랐던 그대와 단 둘이 손 잡고
알 수 없는 이 떨림과 둘이 걸어요 봄바람 휘날리며 흩날
리는 벚꽃 잎이 울려 퍼질 이 거리를 둘이 걸어요 봄바람
휘날리며 흩날리는 벚꽃 잎이 울려 퍼질
이 거리를 둘이 걸어요 그대여 우리 이제 손 잡아요
이 거리에 마침 들려오는 사랑 노래 어떤가요 사랑하는
그대와 단둘이 손잡고 알 수 없는 이 거리를 둘이 걸어요

원제 : 벚꽃엔딩
추천 제목 : 벚꽃 사랑노래

동해물과 백두산이 마르고 닳도록
하느님이 보우하사 우리나라 만세
무궁화 삼천리 화려강산
대한사람 대한으로 길이 보전하세
남산위에 저 소나무 철갑을 두른듯
바람서리 불변함은 우리 기상일세
무궁화 삼천리 화려강산
대한사람 대한으로 길이 보전하세

원제 : 애국가
추천 제목 : 나의 자랑이요 나의 자랑이야

자연스러운 예측 제목 빈도 ↑

- 가사 내용과 예측 제목의 연관성 상승
- 영어를 제거하는 전처리로 한국어로만 이루어진 추천 제목 획득
- 생성된 제목에 '사랑'이라는 단어가 들어가는 편향 문제 일부 개선

5

결론 및 한계점

< 한 계 점 >

- 만족스러운 정답을 얻기 위해 수차례 예측 시도 요구됨
- 한국 정서 : 사랑에 대한 내용 多 → 다양한 주제의 균형된 데이터 추가 학습 필요
- 한국 노래에 영어 가사 상당수 포함, But 전처리로 영어 제외 (제외된 영어 토큰 앞뒤의 문맥 정보 일부 훼손)
  → 다국어 학습 모델 사용 또는 영어 포함 전처리 방안 모색

< 결 론 >

- KoGPT2 Model을 이용해 한글 가사로 노래 제목을 추천해주는 프로그램 구현
- 데이터 추가 및 학습이 진행될수록 예측된 제목과 가사의 연관성이 높아짐 식별 ∴ 데이터 및 학습 ↑
- 데이터 추가 및 학습 증가, 효율적인 데이터 전처리 → 상용화

번 외 추천받은 제목으로 노래 가사 생성

리쌍 - 광대
사회를 살아가는 모든 사람들이
억지로 웃음을 사고 파는 '현대판'
광대라는 의미를 담고 있다.

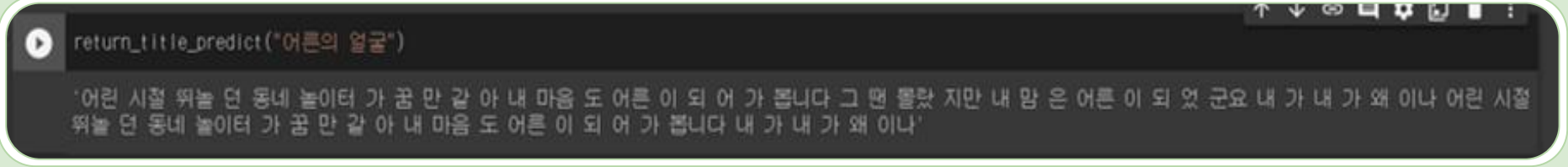
→ 제목

· KoGPT2

어른의 얼굴

→ 가사

?



어린 시절 뛰놀던 동네 놀이터가 꿈만 같아 내 마음도 어른이 되어가 봅니다
그땐 몰랐지만 내 맘은 어른이 되었군요 내가 내가 왜이나
어린 시절 뛰놀던 동네 놀이터가 꿈만 같아 내 마음도 어른이 되어가 봅니다
내가 내가 왜이나

6

팀원 소개 및 역할

김용재

조 장
크롤링 (블로그)
Transformer
디버깅

국승용

크롤링 (블로그)
KoBART
PPT
디버깅

박혜정

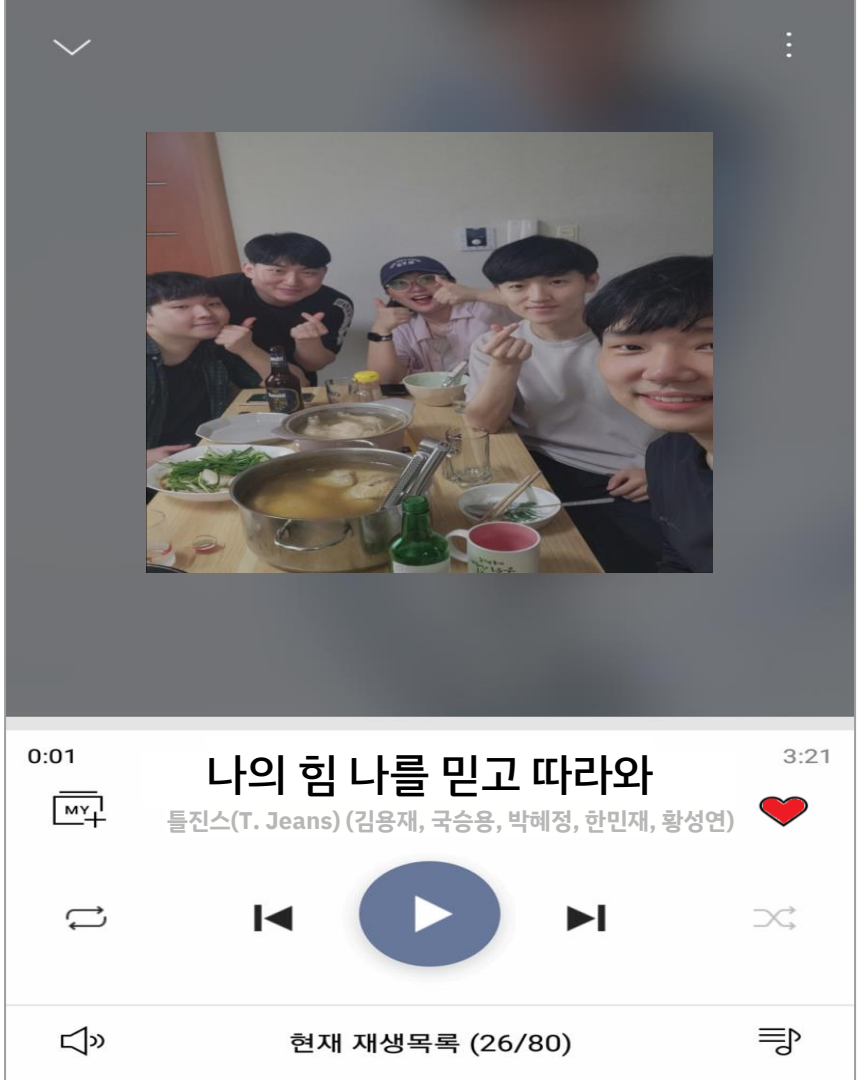
LSTM Attention
크롤링 (멜론)
데이터 전처리
Notion정리
디버깅

한민재

KoGPT2
메인모델 핸들링
코드정리
디버깅

황성연

크롤링 (벅스)
KoBERT
데이터 전처리
디버깅


0:01
나의 힘 나를 믿고 따라와
3:21
틀진스(T. Jeans) (김용재, 국승용, 박혜정, 한민재, 황성연)
현재 재생목록 (26/80)

감사합니다