# 세대별 KPOP 가사 생성 프로젝트

실전 NLP 분반 생성팀
Boy & Girls Generation

금지헌 성유지 이지현 최정윤

# 머리부터 발끝까지....
# 다음으로 생각나는 내용은?


화석

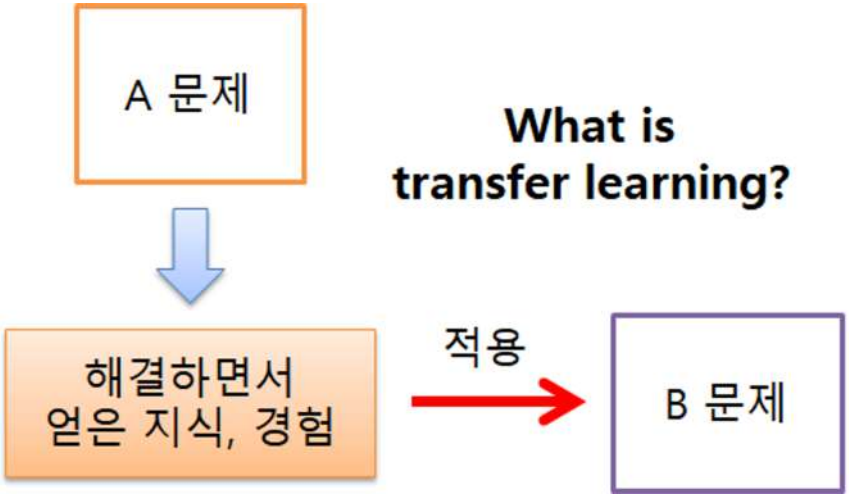

할아버지, 할머니

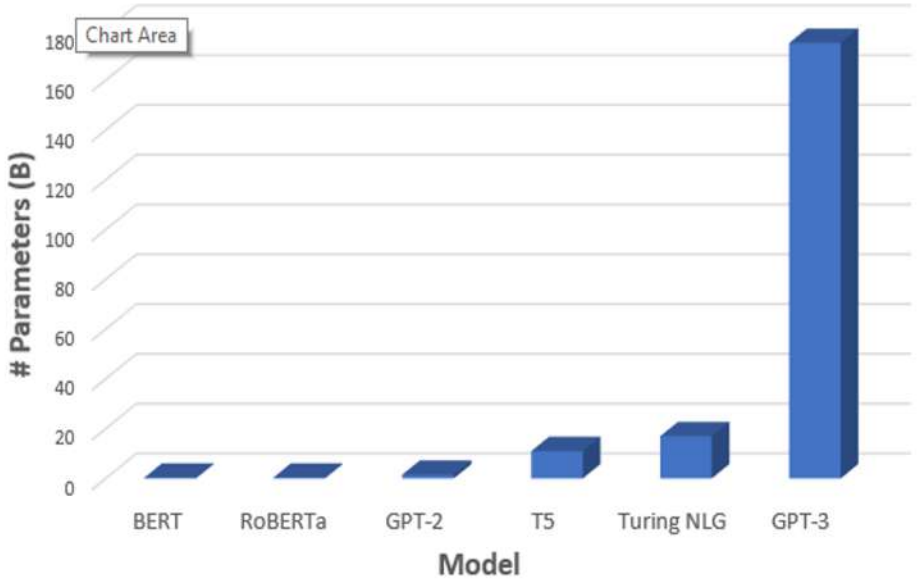

젊은이


요.즘.애.들
"머리부터 발끝까지."

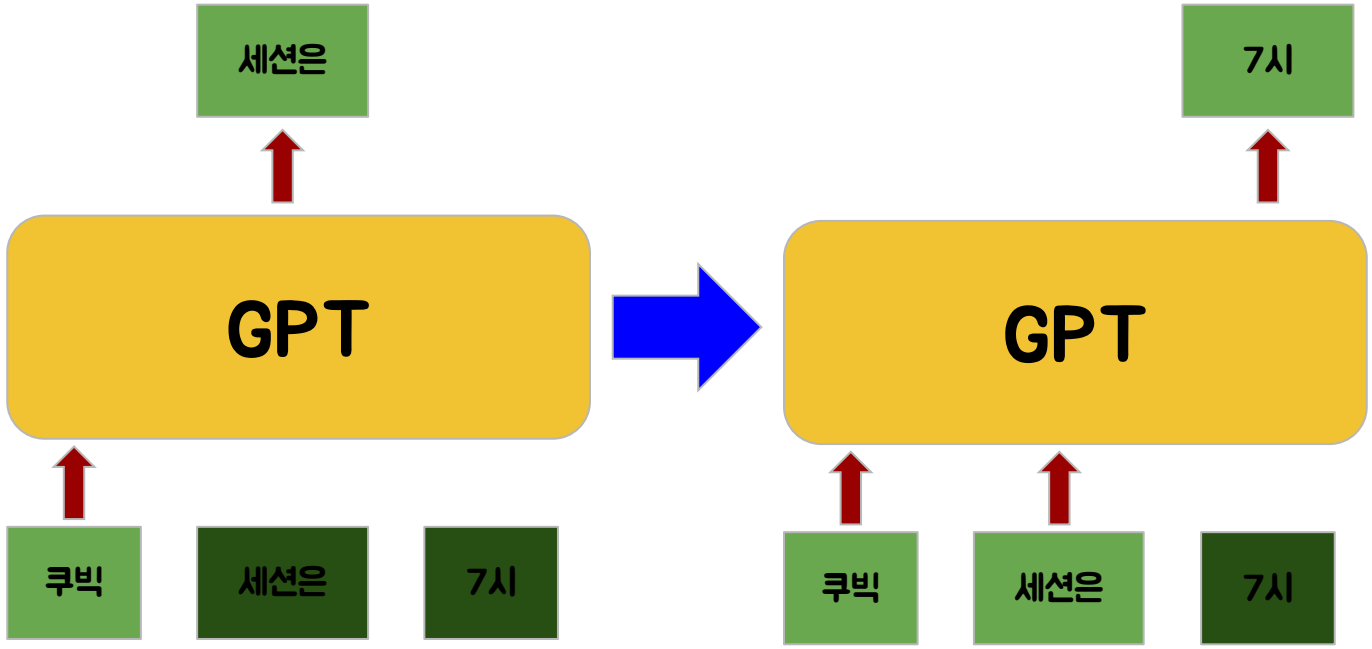
# Pretrained Language Model

# Why Pretrained LM?



# GPT (Self Supervised Learning)

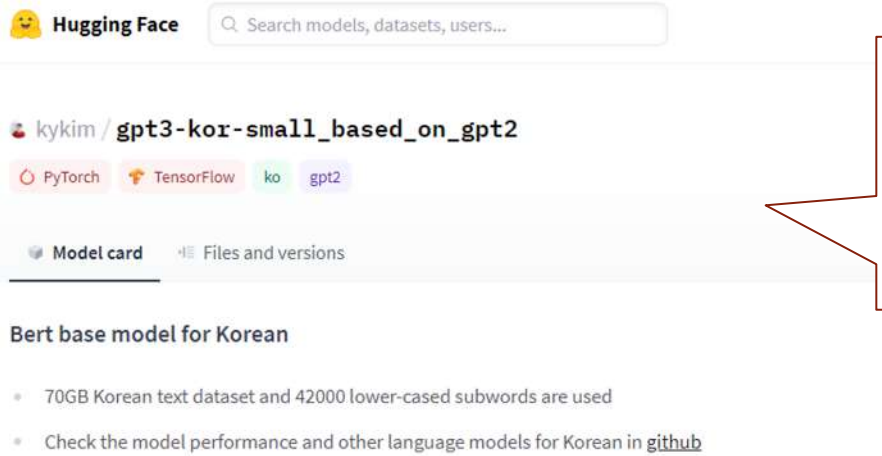


# GPT (Data)

| Dataset | # Tokens (Billions) |
|---|---|
| Total | 499 |
| Common Crawl (filtered by quality) | 410 |
| WebText2 | 19 |
| Books1 | 12 |
| Books2 | 55 |
| Wikipedia | 3 |

약 5,000억 개의 단어 토큰

# Our Model: Pre-trained Language Models For Korean




한국어 데이터를 중심으로 학습한 GPT2 기반 모델, GPT3의 형태 (few-shot learning을 위해 input 길이를 늘리고 계산 효율화 처리)를 반영

- 국내 주요 커머스 리뷰 1억개 + 블로그 형 웹사이트 2000만개 (75GB)
- 모두의 말뭉치 (18GB)
- 위키피디아, 나무위키 (6GB)

최종적으로 70GB (약 127억개의 token) 데이터 학습

# Our Data

| title | artist | gender | lyrics | release_date | debut_year |
|---|---|---|---|---|---|
| The Best | 에이치오티 | 남 | 새로운 세계 속에<br>금지된 세계란 없는 거란<br>생각들로 가득차 있는<br>우... | 2010.07.15 | 1996 |
| I Yah! | 에이치오티 | 남 | 아이야<br>니가 속한 세상에 넌<br>너무너무나도 아름다운<br>세상속에 넌 그렇... | 2002.12.13 | 1996 |
| 그래 그렇게 | 에이치오티 | 남 | 그대로 거기 멈춰선<br>당신의 힘겨운 발걸음을 보았죠<br>세상 앞에 홀로 서기가... | 2002.12.13 | 1996 |
| 아이야 (I Yah!) | 에이치오티 | 남 | 아이야 니가 속한 세상에 넌<br>너무너무나도 아름다운 세상속에 넌<br>그렇게 ... | 2001.05.01 | 1996 |
| 투지 (Git It Up!) + 전사의 후예 + You Got Gun | 에이치오티 | 남 | 모두 다 git it up<br>모두 다 길 잃어 가는<br>모든 우리들의 자신을 ... | 2001.05.01 | 1996 |

✖ 3 (1세대, 2세대, 3세대)

데이터 구성: 노래 제목, 가수명, 가수 성별, 가사, 발매년도, 데뷔년도

# Data Preprocessing

- 데이터프레임 통합 후 'gen' (generation; 세대) 열 생성

| title | artist | gender | lyrics | release_date | debut_year | gen |
|---|---|---|---|---|---|---|
| The Best | 에이치오티 | 남 | 새로운 세계 속에<br>금지된 세계란 없는 거란<br>생각들로 가득차 있는<br>우... | 2010.07.15 | 1996 | 1세대 |
| I Yah! | 에이치오티 | 남 | 아이야<br>니가 속한 세상에 넌<br>너무너무나도 아름다운<br>세상속에 넌 그렇... | 2002.12.13 | 1996 | 1세대 |
| 그래 그렇게 | 에이치오티 | 남 | 그대로 거기 멈춰선<br>당신의 힘겨운 발걸음을 보았죠<br>세상 앞에 홀로 서기가... | 2002.12.13 | 1996 | 1세대 |

| | |
|---|---|
| 1세대 | 2425 |
| 2세대 | 3532 |
| 3세대 | 2636 |

- 데이터 크롤링으로 딸려온 태그들 제거, 띄어쓰기를 의미하는 <br> 태그는 노래 가사를 끊어주는 토큰의 기능을 할 수 있겠다 판단해 남겨놓음

The Best - H.O.T

새로운 세계 속에<br>
금지된 세계란 없는 거란<br>
생각들로 가득차 있는<br>
우리들의 또 다른 반란<br>

어떤 안녕 - 멜로디데이

멀어진다 뜨거워진다 눈물이다<br>
이별이다<br>
그 자리에 주저앉아 눈물이 막<br>
쏟아진다 안녕<br>

# Data Preprocessing

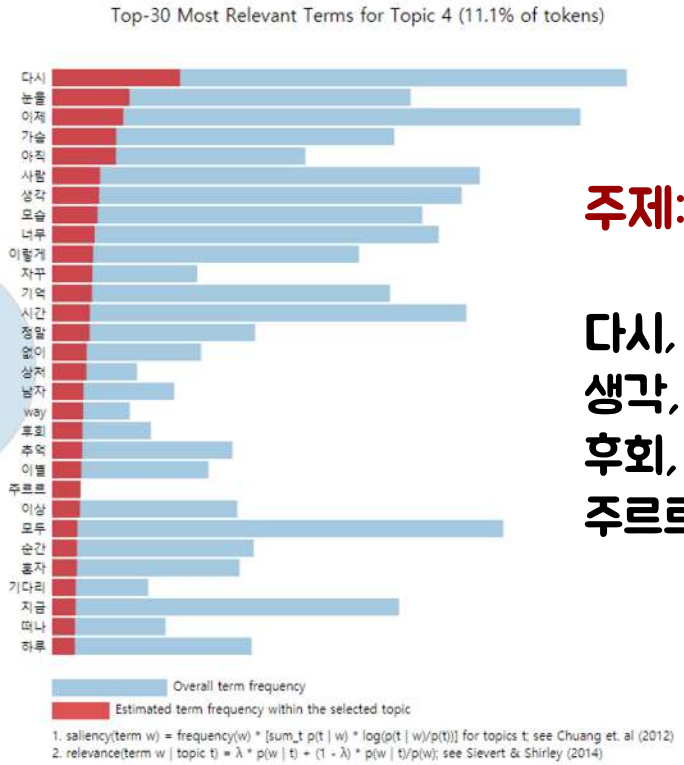
- 아이돌 간 대화 등 노래가 아닌 이상한 데이터 삭제

| title | artist | gender | lyrics | release_date | debut_year | gen |
|---|---|---|---|---|---|---|
| 12:00N | 신화 | 남 | 그녀는 늘 척척 짜여진 스케줄에 맞춰 나를 만났지<br>눈에 맞는 남자를 만나지 못... | 2001.03.29 | 1998 | 1세대 |
| 단장 | 쥬얼리 | 여 | (작곡 : 꼬마 / 작사 : 안영민 / 편곡 : 꼬마 / Guitar : 홍준호 /... | 2005.03.15 | 2001 | 1세대 |
| Closing Ment | 동방신기 | 남 | 윤호 : 어<br><br>재중 : 여러분<br><br>윤호 : 와 ~ 여러분 재밌어... | 2009.07.30 | 2004 | 2세대 |
| Ghost | 히스토리 | 남 | 핏빛 서린 밤 2:00 AM<br>Gonna be stronger<br>판단의 여유... | 2015.05.21 | 2013 | 3세대 |
| SKIT : One night in a strange city | 방탄소년단 | 남 | 지민: 맨날늦어. 내가 한마디 해?<br>진: 그래 지민아 한마디 해.<br>제이홉... | 2015.11.30 | 2013 | 3세대 |
| Skit : R U Happy Now? | 방탄소년단 | 남 | 슈가: 어디야 여기<br>뷔: 잰 차에 타자마자 자냐<br>랩몬스터: 몇 시야 몇 ... | 2013.09.11 | 2013 | 3세대 |
| Skit : Circle Room Talk | 방탄소년단 | 남 | 랩몬스터: 대박이었다니까 그때는 그게<br>제이홉: 2006년<br>랩몬스터: Fl... | 2013.06.12 | 2013 | 3세대 |

- 분석이 힘든 맴버 파트 별로 나눠져 있는 가사, 중국어 가사 등 삭제

| title | artist | gender | lyrics | release_date | debut_year | gen |
|---|---|---|---|---|---|---|
| My Little Princess (상근니설 / 想跟口說) | 동방신기 | 남 | How can I forget all the special memories of y... | 2005.01.18 | 2004 | 2세대 |

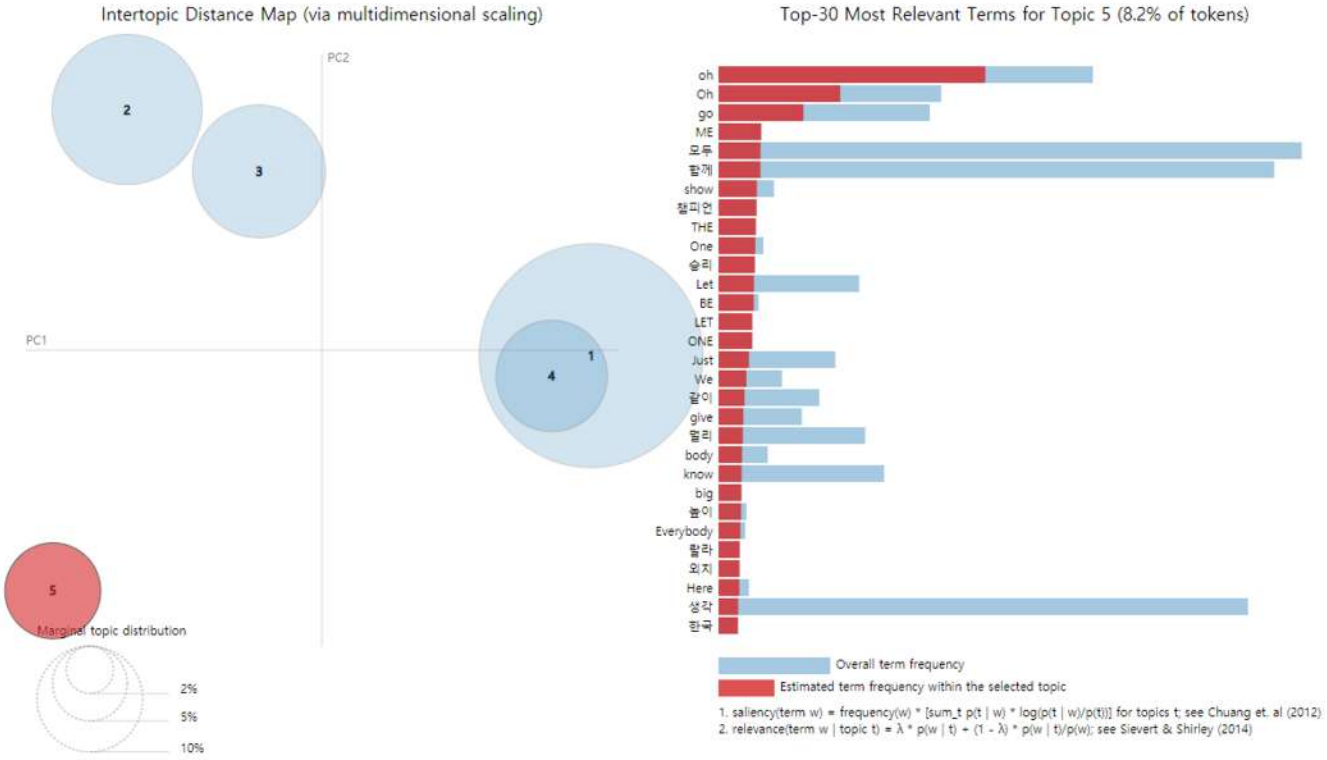
# Topic Modeling - 1세대


**주제: (절절한) 이별**

다시, 눈물, 가슴, 사랑,
생각, 모습, 기억, 추억,
후회, 너무, 상처
주르르, 혼자, 기다리 ...

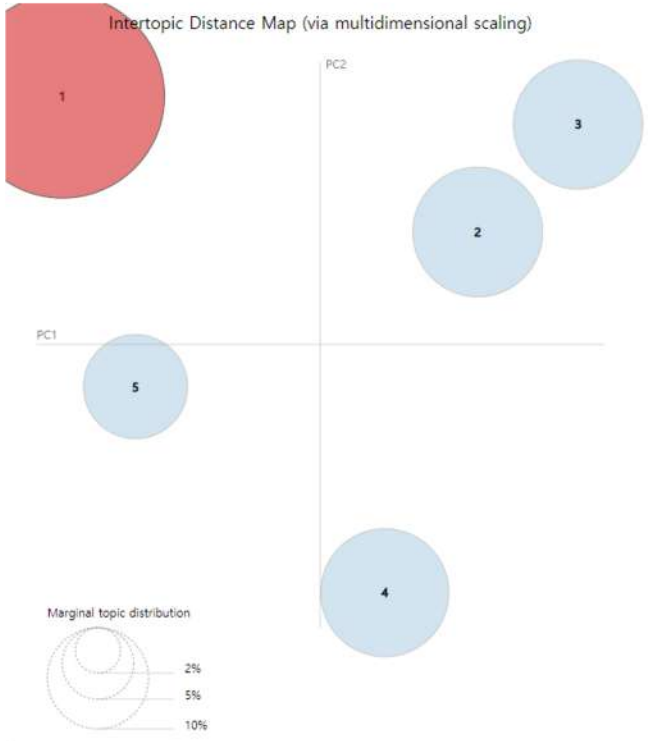
# Topic Modeling - 1세대

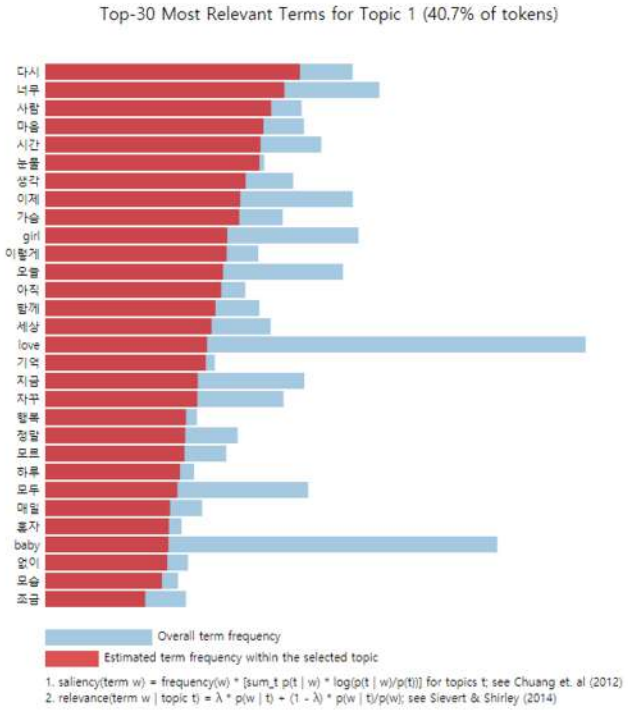



**주제: 챔피언, 승리**

모두, 함께, 같이,
Everybody,승리,
Let, ONE, We,
높이, 외치, 한국 ...

# Topic Modeling - 2세대



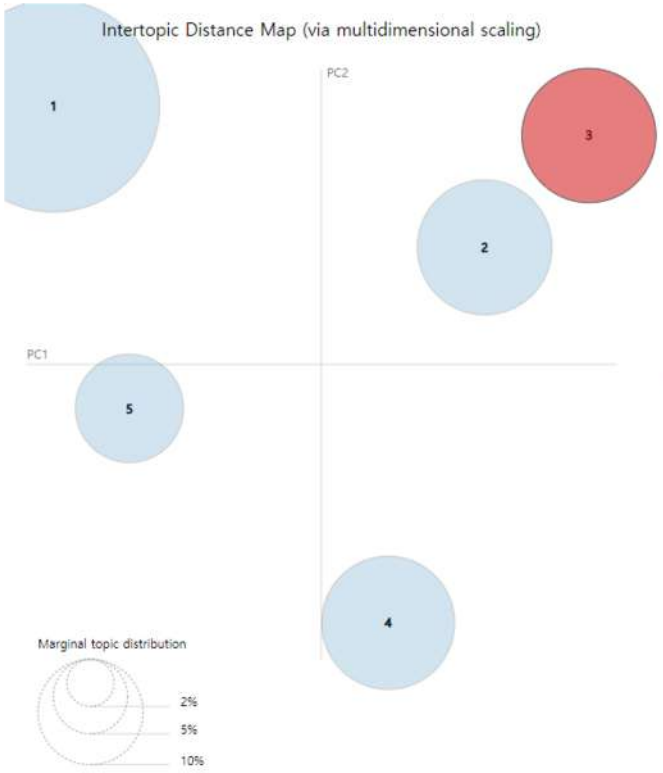



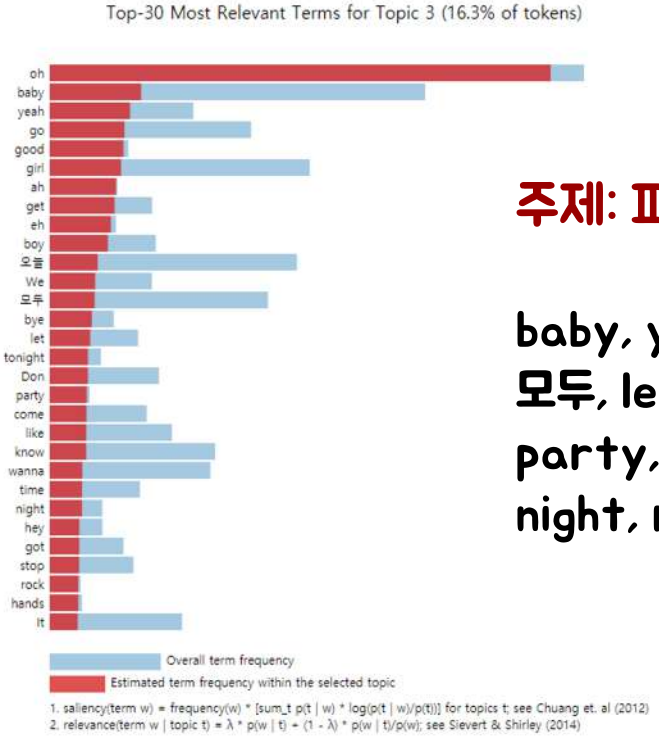

**주제: 사랑, 이별**

사랑, 마음, 시간, 눈물,
생각, 이제, girl, love,
기억, 모르, baby, 조금 ...

➡ 1세대에 비해 덜 무거움

# Topic Modeling - 2세대

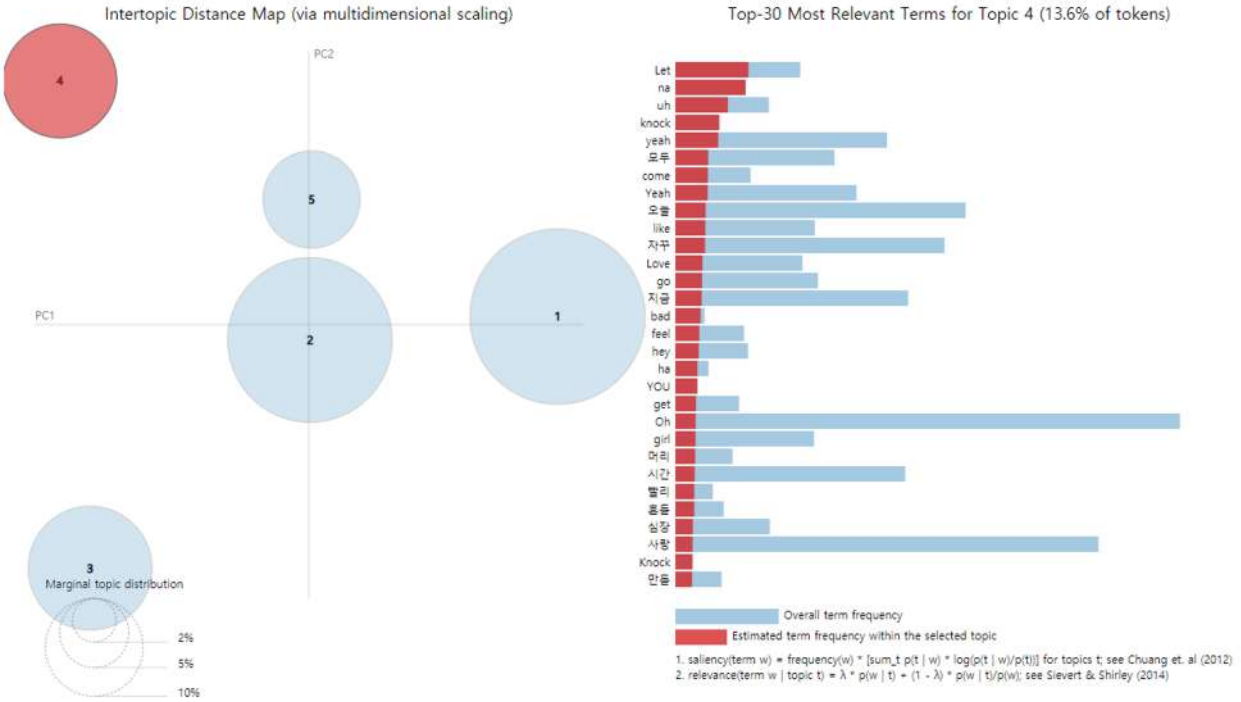



**주제: 파티**

baby, yeah, girl, boy,
모두, let, tonight,
party, wanna, time,
night, rock, hands ...

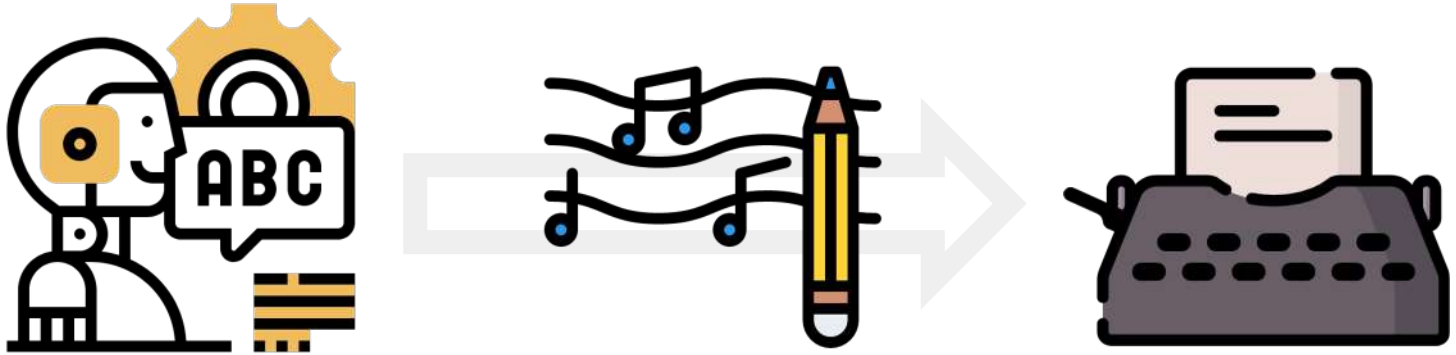
# Topic Modeling - 3세대




**주제: 두근거림**

knock, come, 오늘,
like, 자꾸, Love, feel,
YOU, 머리, 시간, 빨리
흔들, 심장, 사랑 ...

# Fine Tuning




Pre-trained 된 GPT 모델을 우리가 갖고 있는 아이돌 가사 데이터에 맞게 학습시키자!

# Fine Tuning in Our Model

Oh maybe maybe
이건 사랑일지도 몰라
이미 내 눈에는 너만 보여 난
아직 아닐 거라 겁을 먹어도
Oh 내게도 온 것 같아

책에 적힌 글자들이 현실이 돼
가슴이 곧 터져 버릴 것 같아
Oh 아냐 어떤 말로도 설명할 수 없어
그래 나에겐 너뿐이야

**세대, 줄바꿈 토큰 추가**

<1세대> Oh maybe maybe <br>
이건 사랑일지도 몰라 <br>
이미 내 눈에는 너만 보여 난 <br>
아직 아닐 거라 겁을 먹어도 <br>
oh 내게도 온 것 같아 <br>

책에 적힌 글자들이 현실이 돼 <br>
가슴이 곧 터져 버릴 것 같아 <br>
oh 아냐 어떤 말로도 설명할 수 없어 <br>
그래 나에겐 너뿐이야 <br>

**Special Token 추가 후, 8000개 이상의 가사로 Fine tuning 진행!**

# Before Fine Tuning

🎤 입력된 가사: 자 이제 시작이야 내 꿈을

〈FINE TUNING 이전〉

자 이제 시작이야 내 꿈을 네. 오늘 이준영 씨도요.

자 이제 시작이야 내 꿈을 그냥 쭉 따라가다 보면 어 ~ 이런 걸 볼 수 있지 않겠니.

자 이제 시작이야 내 꿈을 내가 인제 그 꿈은

# After Fine Tuning

🎤 입력된 가사: 자 이제 시작이야 내 꿈을

〈FINE TUNING 이후〉

자 이제 시작이야 내 꿈을
언제나 하던대로 나를 하고픈 대로
할 수만 있다면 더 크게 소리쳐봐
저 하늘을 향해 크게 외쳐봐
just one way we come to me
you will be fine and everything
ill never let the rhyme of you girl
날 위해 춤춰 줘 너의 맘을 내가 가질 수 있게
지금 니 앞에 있는 것처럼 느껴줘 함께 해줘

whenever your soul goes around
im ready for survival now
all i feel so tight
그 어떤 말도 필요 없어 널 위해서
오늘 이 밤이 끝날 때까지
내게만 보여 줄 거야 on my mind
넌 잘 모를거야 아마 그럴꺼야
그저 너와 나 사이의 작은 끈으로
끈을 놓지 않을테니까

# FINAL LYRICS

## 🎤 입력된 가사: 머리부터 발끝까지 다

### [1세대 김종국 스타일] 💿 제목: Right

머리부터 발끝까지 다
난 너에게 빠진거야
언제나 가까이에 있어줘
나의 손을 잡아줘
니 옆에 있으면 마음이 놓여
너를 원해 항상 곁에 두고 싶어
내 모든 걸 네게 줄꺼야
영원히 널 위해서만 아껴줄께
이렇게 사랑할 수 있는 날까지

oh baby i love you to me my love forever
wanna be with u babe u right on boy
even though talk about that days and presses
what im true in your heart
눈을 뜨면 내가 꿈꿔왔던 세상속에
내가 존재 한다는 것만으로
너무 감사하고 좋아
매일 너와 함께 할 사랑을 바래 워

# FINAL LYRICS

## 🎤 입력된 가사: 머리부터 발끝까지 다

### [2세대 포미닛 스타일] 💿 제목: Let's

머리부터 발끝까지
lets get it on lets dance
1 2 3 go lets jump it up
모두 손을 위로 들어
우리 모두 다 미쳐
everybody big girls
come to the top 올라 가봐
were rock and holla halla
we drop that roll in this head
오늘밤이 가기 전에 놀아보는 거야

baby baby baby party tonight
now lets break it down let me know that
all night long with your party yo
모두 눈을 뜨고 지금 이 순간을 즐겨
lets get em high ride higher
소리 질러 더 높이 뛰어올라
너와 나를 위한 이 밤의 fever
밤새도록 미친듯이 놀자 ah
리듬에 몸을 맡겨 봐 이제 느껴봐
oh yeah ah oh yeah ah

hot chocolate 보다
달콤한 너의 sugar style
lets shake it slice turn it on
left right right right tonight
get your hands out of my life yo
모두 발을 올리고 지금이 순간만을 즐겨
lets get em high ride higher
소리 질러 더 높이 뛰어올라
너와 나를 위한 이 밤의 fever
밤새도록 미친듯이 놀자

# FINAL LYRICS

## 🎤 입력된 가사: 머리부터 발끝까지 다

### [3세대 아스트로 스타일] 💿 제목: Lover

머리부터 발끝까지
모두 완벽해 너무 예뻐
난 니가 있는 공간에 이렇게 예쁜 girl
이런 흔하지 않은 일이 내 눈앞에 펼쳐져
oh oh yeah 눈이 번쩍 띄지
너의 표정에 정말 반했어 yeah
i know you are perfect man
느낌이 왔어 너란 빛 um
날 부르는 목소리

마치 special lover lover
its all about the way for me girl
that i think im ready to go
with the baby girl this time
내게로 와 다가와 줘요 좀 더 가까이 와요
우리 사랑을 느껴봐요
make u feel higher in white skinny night
숨이 멎는 걸요 나를 보는 그대 눈빛이 말해줘요
lady 달콤한 키스를 해줄래요

kiss 해볼래 kiss her
심장이 쿵쾅쿵쾅 뛰는 소리가 들려
이 떨리는 입술에 입을 맞출까
I need now should've seen
that beat pack started take off tic
to cash don't stop look so good
어서와요 나의 모든걸 get up
니 손짓 하나에도
푹 빠질것만 같아요 babe

# FINAL LYRICS

## 🎤 입력된 가사: 여보세요

### [1세대 백지영 스타일] 💿 제목: 안녕

여보세요
나 조금 늦더라도 괜찮지 않을 거에요
내가 집에 들어갈 때 쯤이면 되요
그 사람이 오는 길이라면 다 달려올께요
혹시 혹시라도 날 돌아봐 줄 것 같아요
다른 사람은 보지 말아요
내 가슴속 깊이 감춰진
너를 찾아낼 수 있으니까
한번만 내게 기회를 줘요

조금만 기다려줘요
사랑하지 않아도 돼
하지만 영원히 널 느낄 순 없어
시간이 가면 갈수록 더 초라해져요
나를 슬프게 해도 좋아 이젠 안녕 안녕
너무 쉽게 다가서는 너의 사랑을 멈출 수는 없잖아
다시는 오지 않아 넌 정말 행복해야 해
한번 두번 세번 네 앞에서 돌아서는 난 바보가 됐어
지금 이대로 널 보낼순 없으니깐

# FINAL LYRICS

## 🎤 입력된 가사: 여보세요

### [2세대 2AM 스타일] 💿 제목: 못했다며

여보세요
잘 지냈어 내가 없이도
그래 웃으며 지내다 왔잖아
우리 함께 했던 시간들이 다 고맙다고
가끔 들리는 너의 전화에
너무 놀라서 말을 못했다며 울던 너
내 맘을 들었다 놨다 하는 사랑스런 너는
지금 무슨 생각하니 oh no please
니가 나를 떠난 후에야 나는 알게 되었지
나의 아픔들 위로해준 너를
이제는 난 이해할 수가 있어

날 잊으려 하지 말아줘 그대 제발요
my love 이젠 알아줬으면 해
더 이상 이대로 끝낼 순 없어
나 정말 미안해 어떻게라도 널 잡을게 ye
your love forever baby
so i cry never trust you
im gonna take it on to the sky
또 다른 사랑이 온다 해도 i cant get enough
your mind just let me show
you answers making me high
내게 남은 상처 모두 지워버리기로 해

네가 없는 예전 모습은 온데간데 없고
언제나 제자리에 항상 서있는 너
언제까지나 그대로 함께 할 거야
나의 마음속 그 곳에 다가가 말하고 싶어
이런 나와는 비교하지 마 제발 돌아와 줘
이 밤이 지나면 넌 없을거야
너무나 그리운 너의 목소리는 듣지 못하게
다시는 돌아올 수 가 없게 될꺼야
영원히 bye bye 이제 나는 괜찮아 u know that
내 곁의 네 모습 잊지 않을게 ooh ooh
think about to lost in time

# FINAL LYRICS

## 🎤 입력된 가사: 여보세요

### [3세대 블랙핑크 스타일] 💿 제목: 따라(DDARRA)

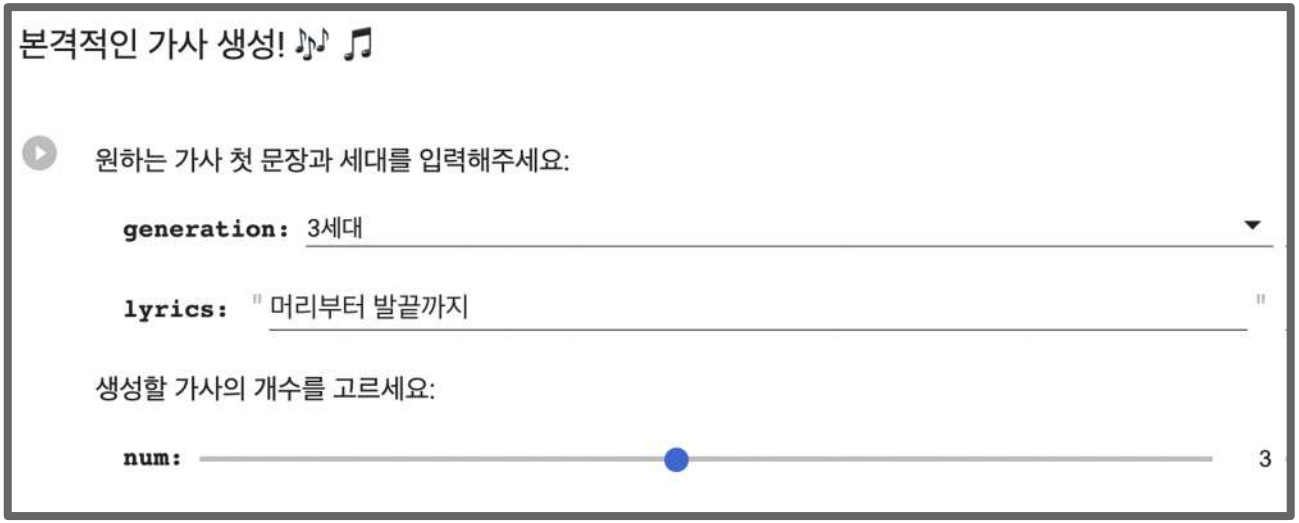
여보세요
자 이리와봐 baby boy now
i wanna get you girls party
gorgeous tonight
날 따라 따라해 봐 ahah
널 마주치고 싶어 너의 향기에 난 취해가 yeah
넌 내게 속삭여 내 귀에 대
hello come on my sweet night
were so beautiful world forever with u

just tell me that im brand new history
우릴 따라 따라 해 너만의 길을 향해
started up baby shout off your door
지금 여긴 너무 뜨거워
더 이상은 못참아 it louder
너를 맡겨둬
오늘밤이 지나면 넌 모두 타버릴거야
oh honey oh honeysome road like this
one more time all right

happen everybody friends
then trend is real really good
what im satistic in seoul yah okay
just tell me whos back in city
니 옆에 있을게 babe 널 기다린다고
이젠 끝이란 말은 안 할거야
우린 여기까지라고 lets pickem high
and we couldnt fight another day
우릴따라 이제 모두 떠나줄께

# MORE LYRICS!

**KUBIG 여러분께서는 어떤 가사를 생성해보고 싶으신가요?**
**내가 원하는 세대와 가사 첫 소절을 알려주세요 😊**



# Song Title

- TfidfVectorizer를 사용하여 주제어 추출
- 단어 빈도-역 문서 빈도 점수를 활용하여 한 가사 안에서 가장 많이 등장하는 단어로 제목을 설정

```
('안녕', 0.2150548503167229),
('달려볼께요', 0.19791544418954266),
('초라해져요', 0.18934587460233518),
('쯤이면', 0.1832656682181894),
('길이라면', 0.178549493749205),
('보낼순', 0.17143809948067618),
('늦더라도', 0.16861589224683618),
('없으니깐', 0.16861589224683618),
('세번', 0.16612652904377442),
('기다려줘요', 0.16389971777785176),
('찾아낼', 0.15835460729463136),
('괜찮지', 0.15533014819064425),
('다가서는', 0.15533014819064425),
('여보세요', 0.15147675307242117),
('행복해야', 0.14723554598829033),
('되요', 0.1429071834852137),
('거에요', 0.14213854753796967),
('들어갈', 0.13931634030412968),
('감춰진', 0.13568425777504647),
('있으니까', 0.13119173787084223),
```

< 1세대 여보세요 가사에 대한 tf-idf 점수 >

# Song Title

여러분께서 생성해주셨던 가사에 제목을 붙여볼까요?

생성된 가사에 제목 붙이기 📝

▶️ ⬅️ click it!

저희 Github의 Colab 링크 방문해주시면
직접 입력해서 실행해보실 수 있습니다!
많은 관심 부탁드려요 ☺️

https://github.com/Lyrics-Generation-Project/Song-Lyrics-Generator

# Limitations & Directions

- 비교적 적은 데이터셋 크기 (8537개)
- 직접 훈련하는 데에는 한계가 있어 사전학습된 모델 사용
- 한국어로 사전학습된 모델이기 때문에 영어 가사의 문맥이 아쉬움
- GPT 이외에도 GAN, 강화학습 등 다른 생성 모델을 사용해볼 수 있음

# 감사합니다 :)