Playlist Recommendation

회귀분석팀 권남택 윤주희 진효주 한유진 황유나

[CONTENTS]

1

주제 선정

- 주제 선정 배경
- 분석 개요

2

데이터 확인

- 데이터 소개
- 데이터 연결

3

데이터 탐색

- 메타데이터
- 학습데이터
- TVT 비교
- LDA

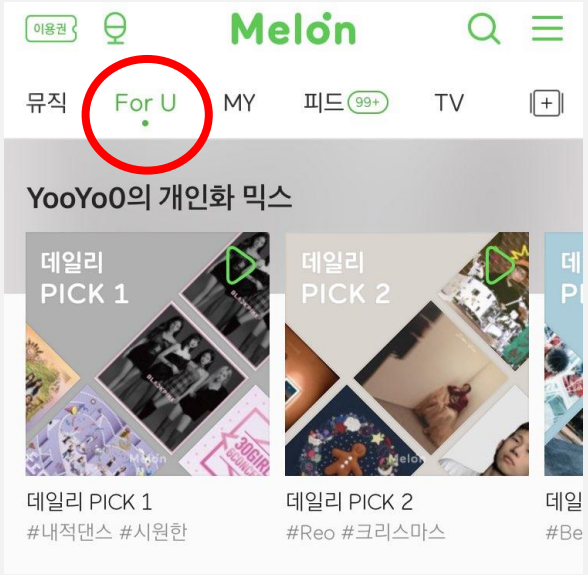
4

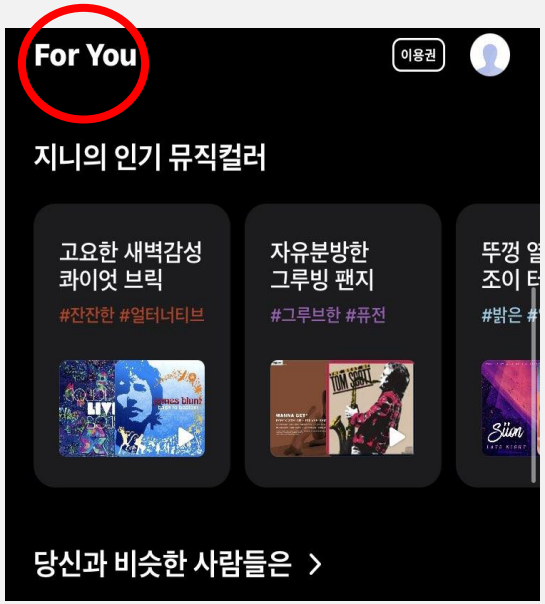
다음주 예고

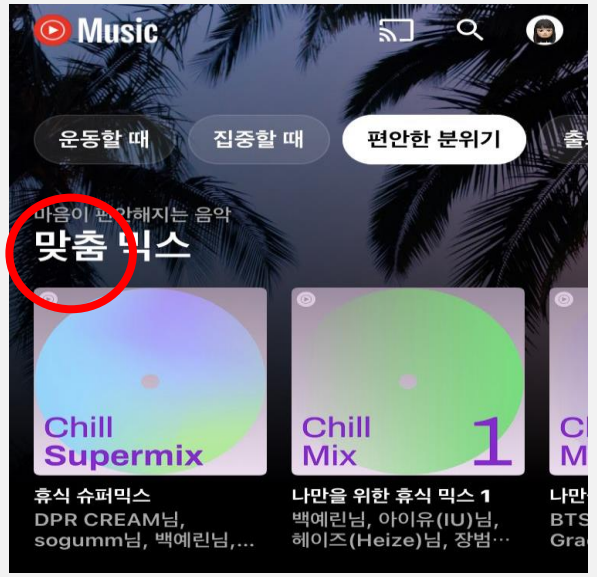
- Word2vec
- 3주차 예고

01. 주제 선정

- 음악 스트리밍 플랫폼







개인화된 추천은 기본적으로 구현되어 있는 기능

01. 주제 선정

- Melon은… 다들 알죠?

음악이 필요한 순간

로고 폰트…충격적…

Melon

대표 뮤직 플랫폼

약 40% 점유율

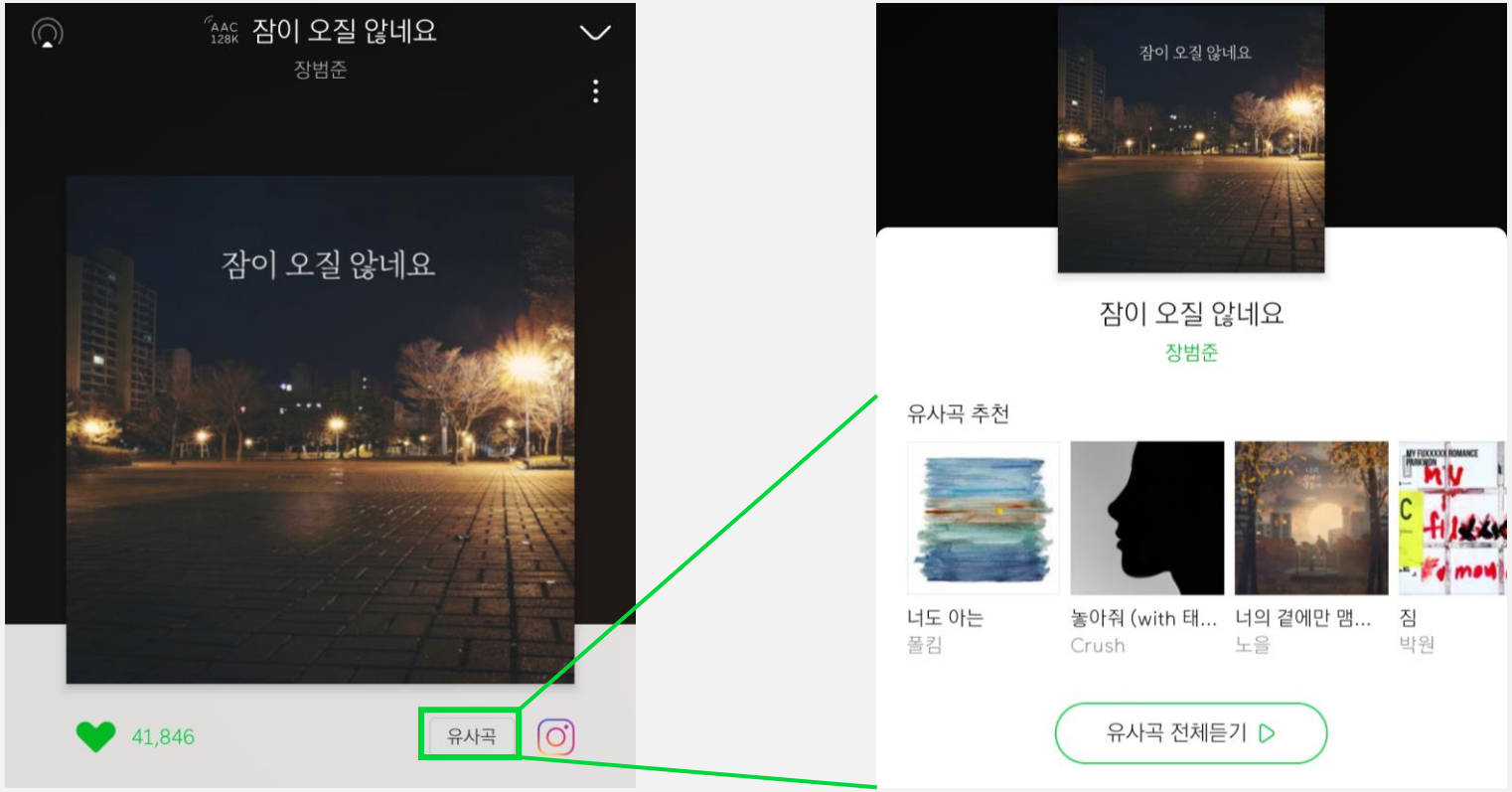
수많은 데이터 기반

취향저격 노래 추천

01. 주제 선정

문제 정의 분석 개요

- Melon의 노래 추천은?

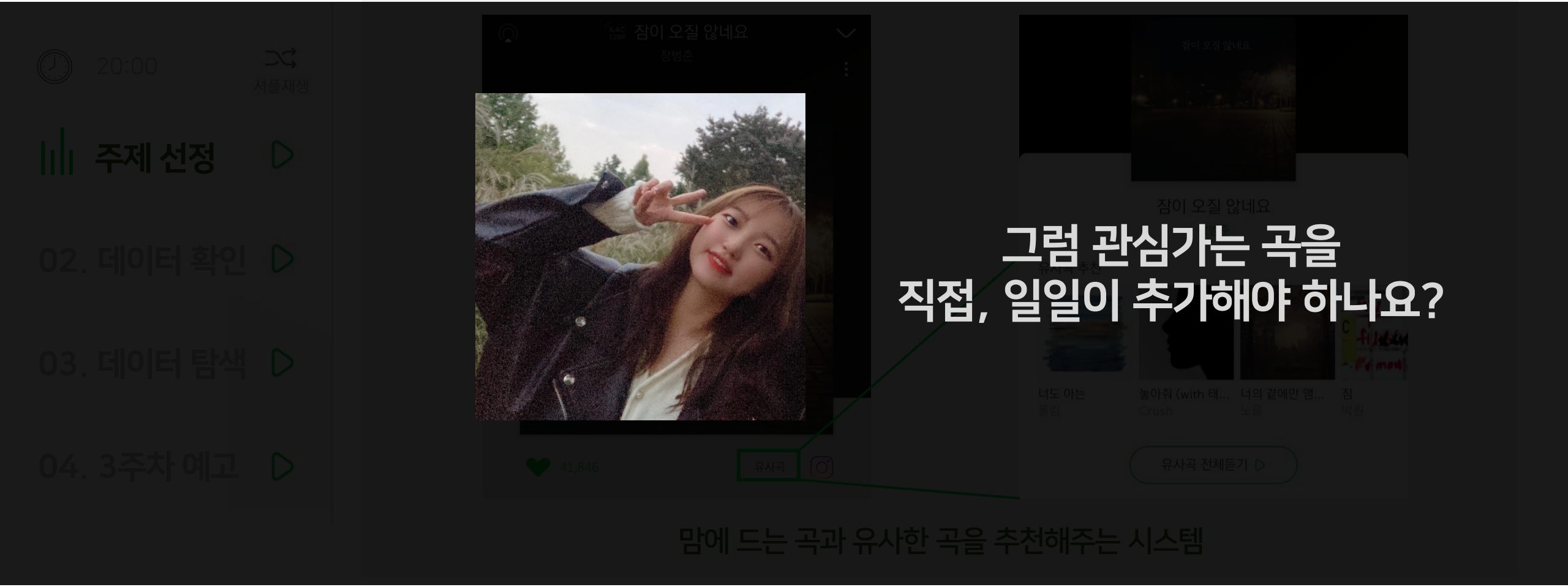


맘에 드는 곡과 유사한 곡을 추천해주는 시스템

01. 주제 선정

문제 정의 분석 개요

- Melon의 노래 추천은?



01. 주제 선정

문제 정의 분석 개요

20:00 셔플재생

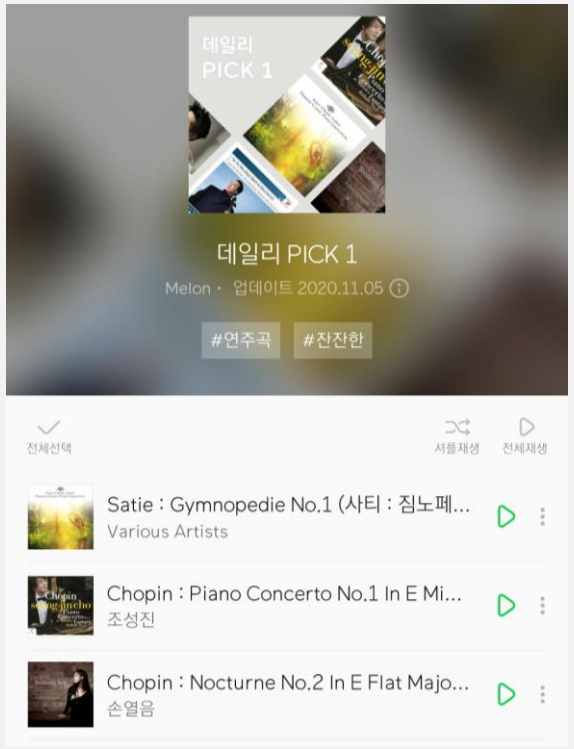
주제 선정

02. 데이터 확인

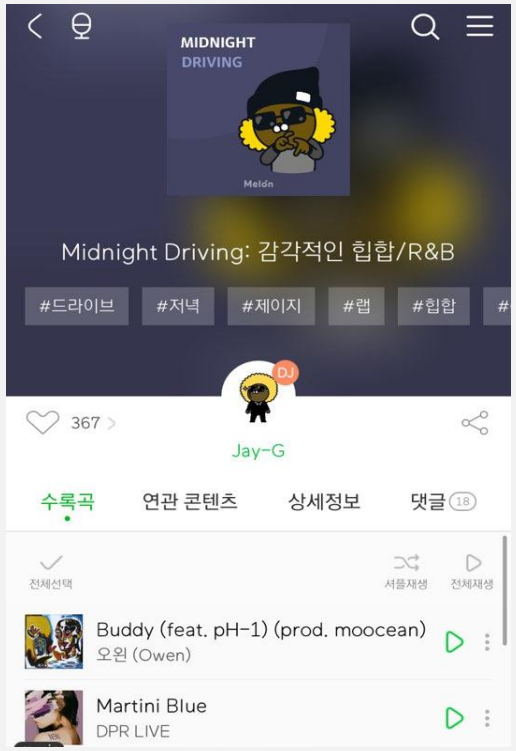
03. 데이터 탐색

04. 3주차 예고

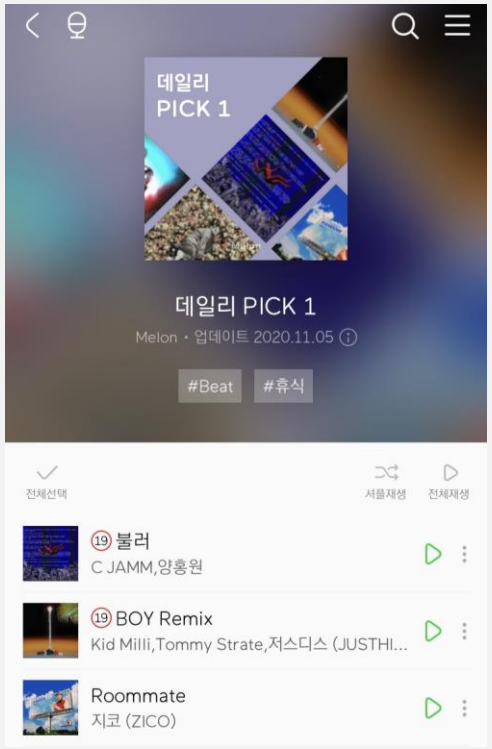
- 다행히 플레이리스트 자체를 추천해준다!



남택 추천



유진 추천

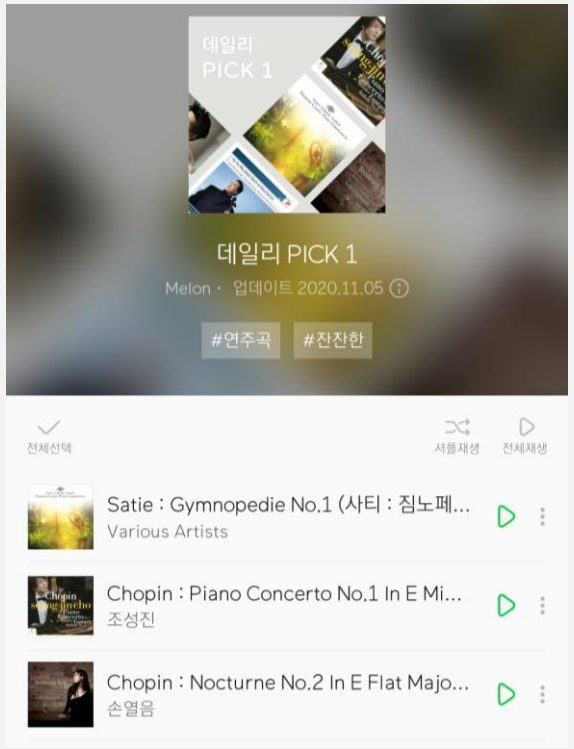


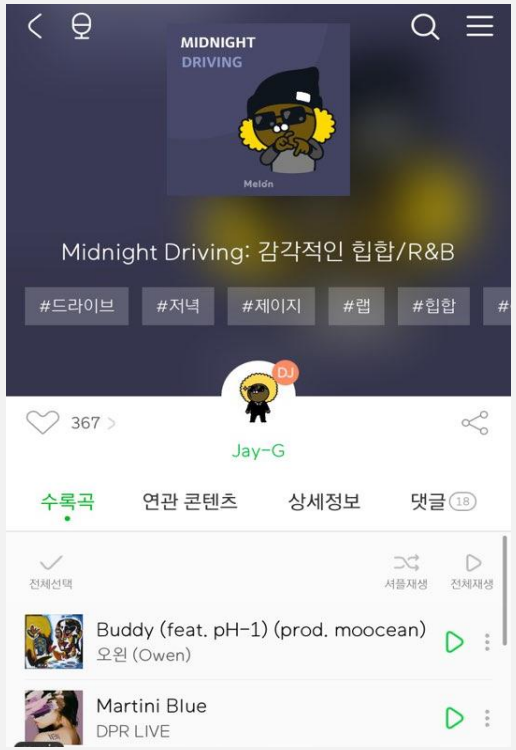
주희 추천

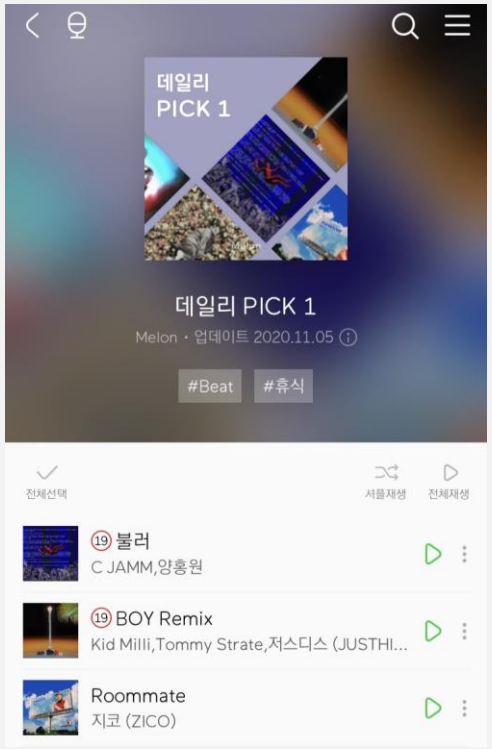
01. 주제 선정

문제 정의 분석 개요

- 정리하자면…







유사한 곡 자체에 대한 추천 + 스트리밍 로그 바탕으로 추천

01. 주제 선정

문제 정의 분석 개요

- 음악 추천이 중요한 이유

Burnt Apple 핵존맛

조주기능사 권씨

Mix & Malt의 칵테일은 14개
돈만 있다면 언제든 다먹을 수 있다…!

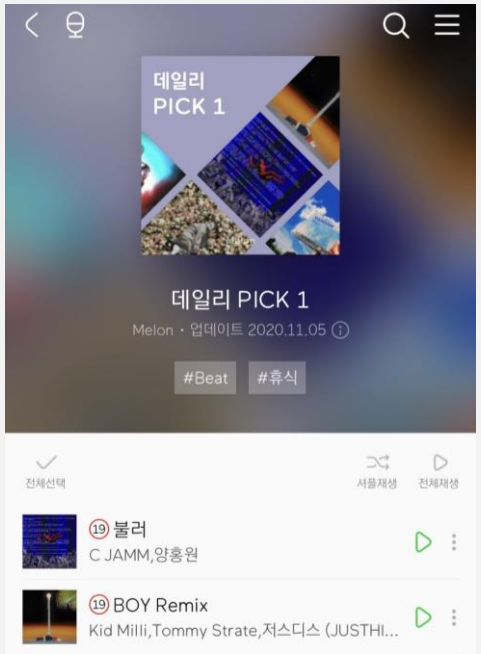
01. 주제 선정

문제 정의 분석 개요

- 음악 추천이 중요한 이유

음악이 필요한 순간

Melon



멜론에서 서비스하는 곡은 수천만…!
모든 곡을 다 들을 수 없다!

01. 주제 선정

문제 정의 분석 개요

- 멜론이 제공하는 플레이리스트 추천

믿고 듣는 세심한 음악 추천, 장르 & 멜론 DJ

세분화된 장르별 플레이리스트와 파워DJ들이 엄선한
고퀄리티 플레이리스트가 매일매일 업데이트돼요.

매일매일 음악 소확행 For U

내 기분과 내가 들은 음악에 따라 맞춤 추천되는, 나를 위한
플레이리스트 'For U 데일리 믹스'를 매일 선물 받아요.

스마트한 주제별 검색

비오는 밤에 어울리는 노래부터, 추억을 소환하는 노래들까지!
궁금한 모든 노래를 검색해보세요.

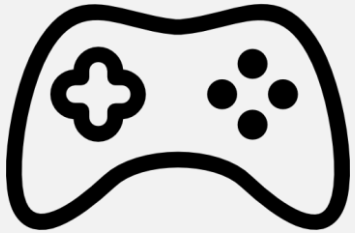
멜론 DJ 플레이리스트

- 누구나 참여 가능
- TPO(Time, Place, Occasion) 고려
- 누적 DJ플레이리스트는 가파르게 증가
- 많은 이용자들이 플레이리스트로 음악 재생

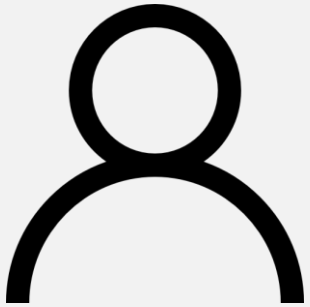
01. 주제 선정

문제 정의 분석 개요

- 멜론이 제공하는 플레이리스트 추천



게임 로그



유저 로그

쇼핑록 품목

시간에 따라 변해서 매번 새로운 모델을 만드는 비용 존재

01. 주제 선정

문제 정의 분석 개요

- 멜론이 제공하는 플레이리스트 추천

song

Playlist

Database

곡과 플레이리스트는 저장되기 때문에 시간에 민감하게 변하지 않음!

Like Netflix…!

01. 주제 선정

문제 정의 분석 개요

- 멜론이 제공하는 플레이리스트 추천

주분 위해 멜론 결제함!!

멜론 DJ 플레이리스트

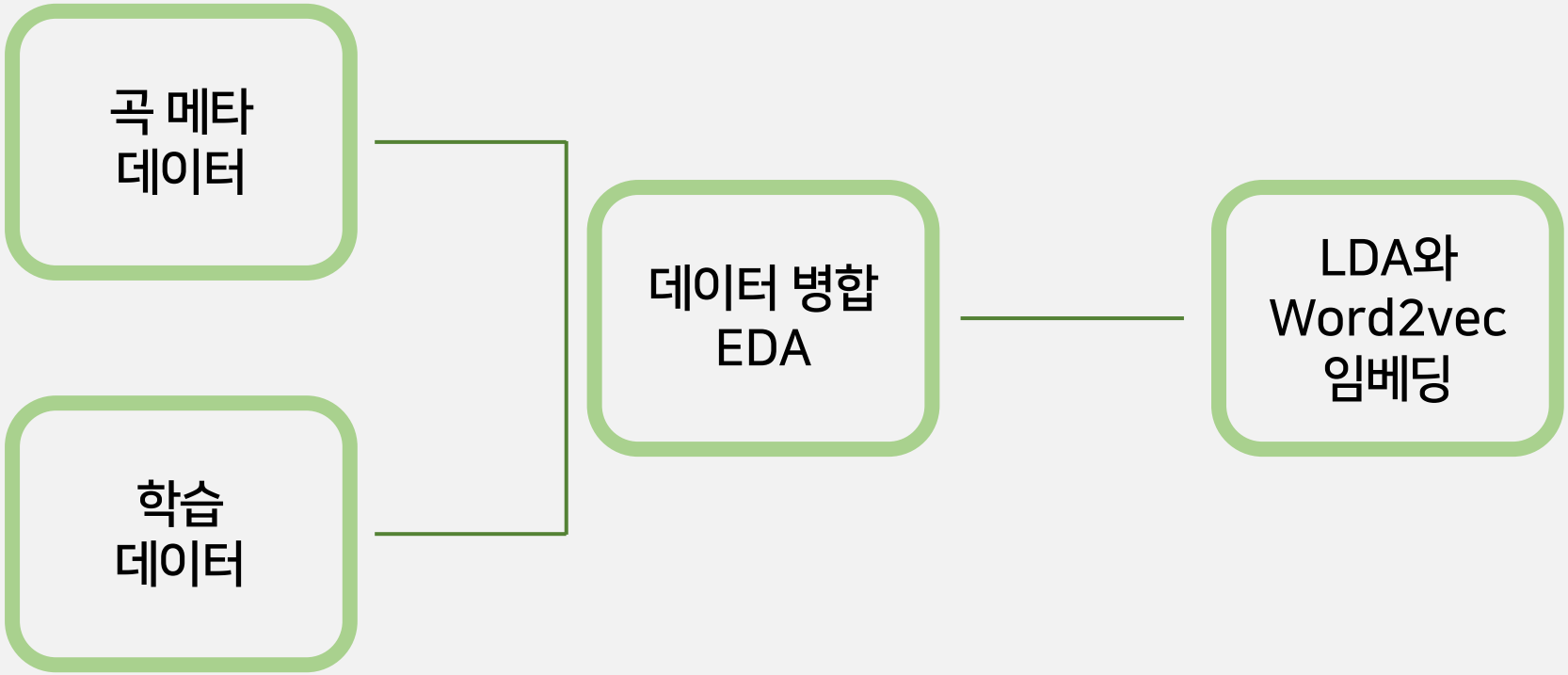
- 누구나 참여 가능
- TPO(Time, Place, Occasion) 고려
- 누적 DJ플레이리스트는 가파르게 증가
- 많은 이용자들이 플레이리스트로 음악 재생

플레이리스트 추천을 고도화 해보자!

01. 주제 선정

문제 정의 분석 개요

- 1주차 분석 방향



02. 데이터 확인

데이터 소개 데이터 연결

genre_gn_all.json

	곡 장르코드
1	gnr_code : 장르 고유 코드
2	gnr_name : 장르명

train/val/test.json

	플레이리스트
1	tags
2	id
3	songs
4	like_cnt
5	updt_date

song_meta.json

	곡 별 메타데이터
1	song_gn_dtl_gnr_basket
2	issue_date
3	album_name
4	album_id
5	artist_id_basket
6	song_name
7	song_gn_gnr_basket
8	artist_name_basket
9	Id

02. 데이터 확인

데이터 소개 데이터 연결

① 장르 구분 코드 데이터 (genre_gn_all.json)

곡 장르코드	
1	gnr_code : 장르 고유 코드
2	gnr_name : 장르명

254행 2열

	gnr_code	gnr_name
0	GN0100	발라드
1	GN0101	세부장르전체
2	GN0102	'80
3	GN0103	'90
4	GN0104	'00
...	...	...
249	GN2900	뮤지컬
250	GN2901	세부장르전체
251	GN2902	국내뮤지컬
252	GN2903	국외뮤지컬
253	GN3000	크리스마스

02. 데이터 확인

데이터 소개 데이터 연결

② 곡 별 메타데이터 (song_meta.json)

세부장르	발매일	앨범명	앨범_id	아티스트_id	곡 명	장르	아티스트_id	곡_id
['GN0901']	20140512	불후의 명곡 - 7080 추억의 얄개시대 팝송 베스트	2255639	[2727]	Feelings	['GN0900']	['Various Artists']	0
['GN1601', 'GN1606']	20080421	Bach : Partitas Nos. 2, 3 & 4	376431	[29966]	Bach : Partita No. 4 In D Major, BWV 828 - I	['GN1600']	['Murray Perahia']	1
['GN0105', 'GN0101']	20160120	행보 2015 윤종신 / 작사가 윤종신 Live Part.1	2662866	[437]	스치듯 안녕	['GN0100']	['윤종신']	707986
['GN1807', 'GN1801']	20131217	명상의 시간을 위한 뉴에이지 음악	2221722	[729868]	숲의 빛	['GN1800']	['Nature Piano']	707987
['GN0601', 'GN0604']	19980000	김경호 Live	34663	[895]	Queen 명곡 멜로디	['GN0600']	['김경호']	707988

707989행 9열

02. 데이터 확인

데이터 소개 데이터 연결

③ train 데이터 (train.json)

태그	playlist id	playlist name	수록 곡 id	좋아요 개수	업데이트 일자
['락']	61281	여행같은 음악	[525514, 129701, 383374, 562083, 297861, 13954...	71	2013-12-19 18:36:19.000
['추억', '회상']	10532	요즘 너 말야	[432406, 675945, 497066, 120377, 389529, 24427...	1	2014-12-02 16:19:42.000
['까페', '잔잔한']	76951	편하게, 잔잔하게 들을 수 있는 곡.-	[83116, 276692, 166267, 186301, 354465, 256598...	17	2017-08-28 07:09:34.000
['잔잔한', '버스', '퇴근버스', 'Pop', '풍경', '퇴근길']	131982	퇴근 버스에서 편히 들으면서 하루를 마무리하기에 좋은 POP	[533534, 608114, 343608, 417140, 609009, 30217...	4	2019-10-25 23:40:42.000
['노래추천', '팝송추천', '팝송', '팝송모음']	100389	FAVORITE POPSONG!!!	[26008, 456354, 324105, 89871, 135272, 143548,...	17	2020-04-18 20:35:06.000

115071행 6열

02. 데이터 확인

데이터 소개 데이터 연결

④ val/test 데이터 (val/test.json)

태그	playlist id	playlist name	수록 곡 id	좋아요 개수	업데이트 일자
[]	118598		[373313, 151080, 275346, 696876, 165237, 52593...	1675	2019-05-27 14:14:33.000
[]	131447	앨리스테이블	[]	1	2014-07-16 15:24:24.000
[]	51464		[529437, 516103, 360067, 705713, 226062, 37089...	62	2008-06-21 23:26:22.00
['잔잔한']	101722		[75842, 26083, 244183, 684715, 500593, 508608,...	17	2015-12-17 14:06:05.000
['어머니', '힘들때', '아빠', '가족', '위로받고싶을때']	122127		[450275, 487671, 561031, 663944, 628672, 59121...	10	2020-04-16 21:35:44.000

Val : 23015행 6열
test : 10740행 6열

02. 데이터 확인

데이터 소개 데이터 연결

④ val/test 데이터 (val/test.json)

태그	playlist id	playlist name	수록 곡 id	좋아요 개수	업데이트 일자
[]	118598		[373313, 151080, 275346, 696876, 165237, 52593...	1675	2019-05-27 14:14:33.000
[]	131447	앨리스테이블	[]	1	2014-07-16 15:24:24.000
[]	64		[529437, 516103, 360067, 705713, 226062, 37089...	62	2008-06-21 23:26:22.00
['잔잔한']	101722		[75842, 26083, 244183, 684715, 500593, 508608,...	17	2015-12-17 14:06:05.000
['어머니', '힘들때', '아빠', '가족', '위로받고싶을때']	122127		[450275, 487671, 561031, 663944, 628672, 59121...	10	2020-04-16 21:35:44.000

Val : 23015행 6열
test : 10740행 6열

플레이리스트 일부의 정보를 이용해

태그, 수록곡을 예측하는 문제!

02. 데이터 확인

데이터 소개 데이터 연결

④ val/test 데이터 (val/test.json)

태그	playlist id	playlist name	수록 곡 id	좋아요 개수	업데이트 일자
	118598		[373313, 151080, 275346, 696876, 165237, 52593...	1675	2019-05-27 14:14:33.000
	131447	앨리스테이블	[]	1	2014-07-16 15:24:24.000
	1464		[529437, 516103, 360067, ...	62	2008-06-21 23:26:22.00
			[...	17	2015-12-17 14:06:05.000
['어머니', '힘들때', '아빠', '가족', '위로받고싶을때']	122127	돌아온 20-1 시계열 귀요미	[450275, 487671, 561031, 663944, 628672, 59121...	10	2020-04-16 21:35:44.000

Val : 23015행 6열
test : 10740행 6열

EDA하기 전에…

데이터를 어떻게
연결시킬 수 있을지 볼까?

02. 데이터 확인

데이터 소개 데이터 연결

- 장르 구분 코드 데이터 (genre_gn_all.json)

	gnr_code	gnr_name
0	GN0100	발라드
1	GN0101	세부장르전체
2	GN0102	'80
3	GN0103	'90
4	GN0104	'00
...	...	...
249	GN2900	뮤지컬
250	GN2901	세부장르전체
251	GN2902	국내뮤지컬
252	GN2903	국외뮤지컬
253	GN3000	크리스마스

대분류 장르와 소분류 장르가
혼재되어 있음

02. 데이터 확인

데이터 소개 데이터 연결

- 장르 구분 코드 데이터 (genre_gn_all.json)

	gnr_code	gnr_name
0	GN0100	발라드
1	GN0101	세부장르전체
2	GN0102	'80
3	GN0103	'90
4	GN0104	'00
...	...	...
249	GN2900	뮤지컬
250	GN2901	세부장르전체
251	GN2902	국내뮤지컬
252	GN2903	국외뮤지컬
253	GN3000	크리스마스

대분류 장르와 소분류 장르가
혼재되어 있음

대분류 코드	대분류 장르	소분류 코드	소분류장르
GN0100	발라드	GN0101	세부장르전체
GN0100	발라드	GN0102	'80
GN0100	발라드	GN0103	'90
GN0100	발라드	GN0104	'00
GN0100	발라드	GN0105	'10-
GN2900	뮤지컬	GN2901	세부장르전체
GN2900	뮤지컬	GN2902	국내뮤지컬
GN2900	뮤지컬	GN2903	국외뮤지컬
GN3000	크리스마스	GN3000	크리스마스

대분류와 소분류를 매칭

02. 데이터 확인

데이터 소개 데이터 연결

- train데이터에 곡 메타데이터 조인

song_meta.json

아티스트_id	곡 명	장르	아티스트_id	곡_id
[2727]	Feelings	['GN0900]	['Various Artists']	0
[29966]	Bach : Partita No. 4 In D Major, BWV 828	['GN1600]	['Murray Perahia']	1
[437]	스치듯 안녕	['GN0100]	['윤종신']	707986
[729868]	숲의 빛	['GN1800]	['Nature Piano']	707987
[895]	Queen 명곡 멜로디	['GN0600]	['김경호']	707988

train.json

playlist name	수록 곡 id	좋아요 개수
여행같은 음악	[525514, 129701, 383374, 562083, 297861, 13954...	71
요즘 너 말야	[432406, 675945, 497066, 120377, 389529, 24427...	1
편하게, 잔잔하게 들을 수 있는 곡.-	[83116, 276692, 166267, 186301, 354465, 256598...	17
퇴근 버스에서 편히 들으면서 하루를 마무리하기에 좋은 POP	[533534, 608114, 343608, 417140, 609009, 30217...	4
FAVORITE POPSONG!!!	[26008, 456354, 324105, 89871, 135272, 143548,...	17

02. 데이터 확인

데이터 소개 데이터 연결

- train데이터

태그	playlist id	playlist name	수록 곡 id	좋아요 개수	업데이트 일자
['락']	61281	여행같은 음악	[525514, 129701, 383374, 562083, 297861, 13954...	71	2013-12-19 18:36:19.000
['추억', '회상']	10532	요즘 너 말야	[432406, 675945, 497066, 120377, 389529, 24427...	1	2014-12-02 16:19:42.000
['까페', '잔잔한']	76951	편하게, 잔잔하게 들을 수 있는 곡.-	[83116, 276692, 166267, 186301, 354465, 256598...	17	2017-08-28 07:09:34.000
['잔잔한', '버스', '퇴근버스', 'Pop', '풍경', '퇴근길']	131982	퇴근 버스에서 편히 들으면서 하루를 마무리하기에 좋은 POP	[533534, 608114, 343608, 417140, 609009, 30217...	4	2019-10-25 23:40:42.000

Train 데이터가 'tidy'한 형태가 아님을 확인 가능!

02. 데이터 확인

데이터 소개 데이터 연결

- Tidy data

  ✓ 각 변수는 개별적인 열(column)으로 존재한다.

  ✓ 각 관측치는 행(row)를 구성한다.

  ✓ 한 관찰 유형은 하나의 테이블을 형성한다.

태그	playlist id	playlist name	수록 곡 id	좋아요 개수	업데이트 일자
['락']	61281	여행같은 음악	[525514, 129701, 383374, 562083, 297861, 13954...	71	2013-12-19 18:36:19.000
['추억', '회상']	10532	요즘 너 말야	[432406, 675945, 497066, 120377, 389529, 24427...	1	2014-12-02 16:19:42.000

분석하기 편하도록 tidy한 형태로 풀어주자!

02. 데이터 확인

데이터 소개 데이터 연결

- train데이터를 tidy 데이터로 – 수록 곡

playlist id	수록 곡 id
61281	[525514, 129701, 383374, 562083, 297861, 13954...
10532	[432406, 675945, 497066, 120377, 389529, 24427...
76951	[83116, 276692, 166267, 186301, 354465, 256598...
131982	[533534, 608114, 343608, 417140, 609009, 30217...
100389	[26008, 456354, 324105, 89871, 135272, 143548,...

playlist id	수록 곡 id
10532	432406
10532	675945
10532	497066
10532	120377
10532	389529

02. 데이터 확인

데이터 소개 데이터 연결

- train데이터를 tidy 데이터로 – 수록 곡

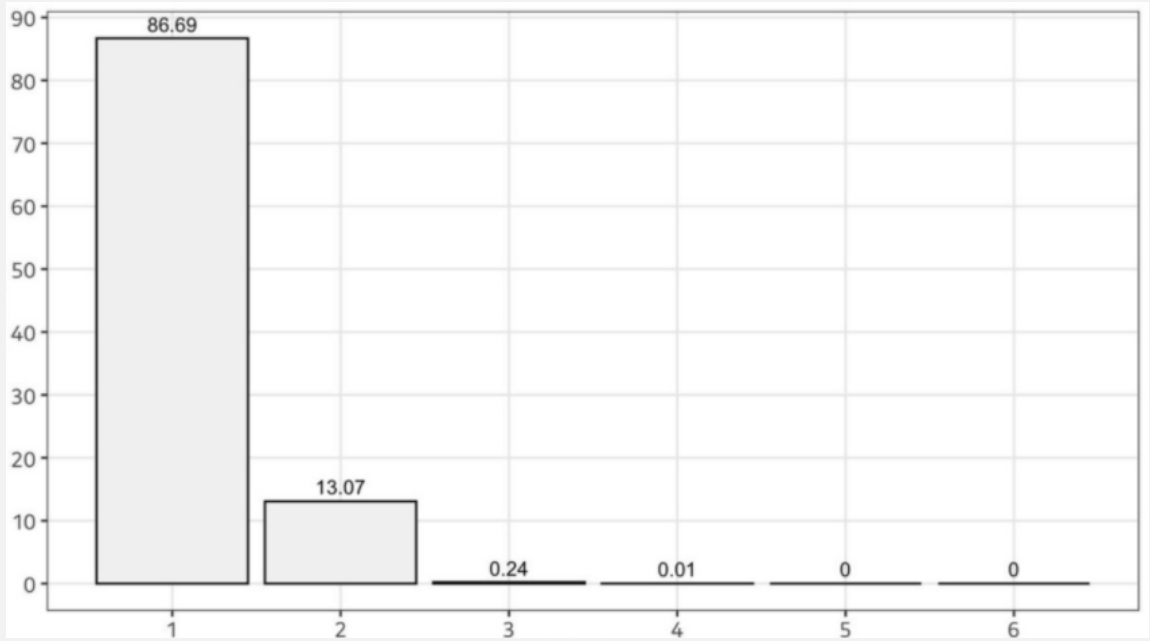
playlist id	태그
61281	['락']
10532	['추억', '회상']
76951	['까페', '잔잔한']
131982	['잔잔한', '버스', '퇴근버스', 'Pop', '풍경', '퇴근길']
100389	['노래추천', '팝송추천', '팝송', '팝송모음']

playlist id	태그
131982	잔잔한
131982	버스
131982	퇴근버스
131982	Pop
131982	풍경

03. 데이터 탐색

- 곡 메타데이터(한 곡이 포함하는 장르 수)

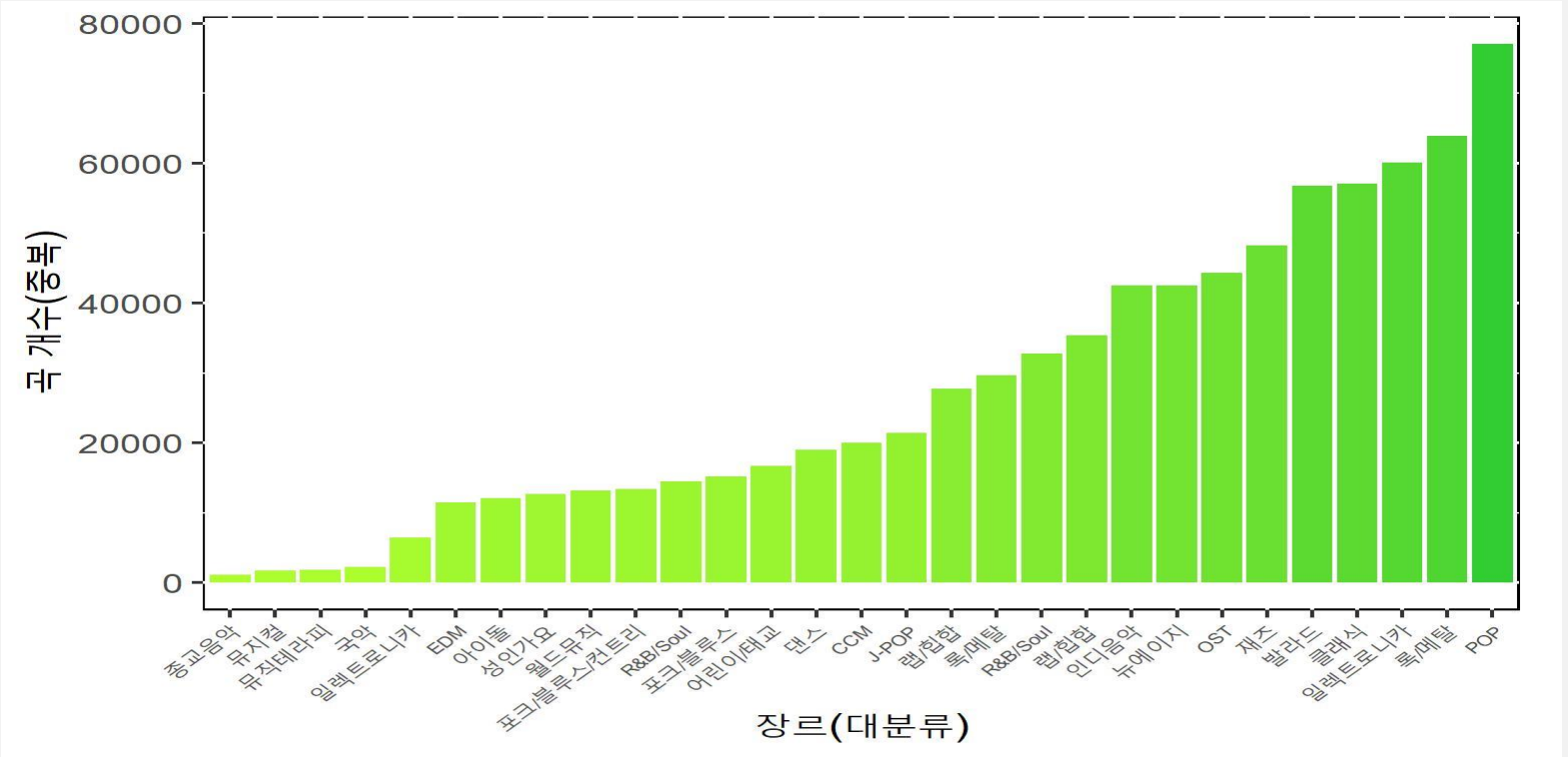
장르 개수	곡 수	비율
1	612806	86.686
2	92378	13.067
3	1694	0.24
4 이상	52	0.007



대체로 노래 한 곡에 한 개의 장르가 할당되지만
약13%의 비율로 2개 이상의 장르를 가진다는 것을 확인할 수 있었다!

03. 데이터 탐색

- 메타데이터 장르별 분류(대분류)



Pop/클래식/발라드 장르가 많았고,
아이돌 곡은 예상보다 인기가 많지 않았다.

03. 데이터 탐색

- '좋아요 개수' 가 많은 플레이리스트의 이름

플레이리스트 이름	좋아요 개수	업데이트 날짜
HOT TRENDY POP: 놓쳐선 안될 'POP' (매주 업데이트)	53211	2020-04-17
♡때껄룩님 TAKE A LOOK 플레이리스트♡	41844	2020-04-23
듣다보면 '우와!' 하고 제목을 보게되는 팝	27268	2020-04-19
약속 있어? 외출 전, 기분 UP 하고 싶을 때 들어봐! [매주 목요일]	23965	2020-04-22
감성이 터지는 팝음악들	23732	2015-07-29

플레이리스트 이름부터 어필하려는 노력이 보임!!

03. 데이터 탐색

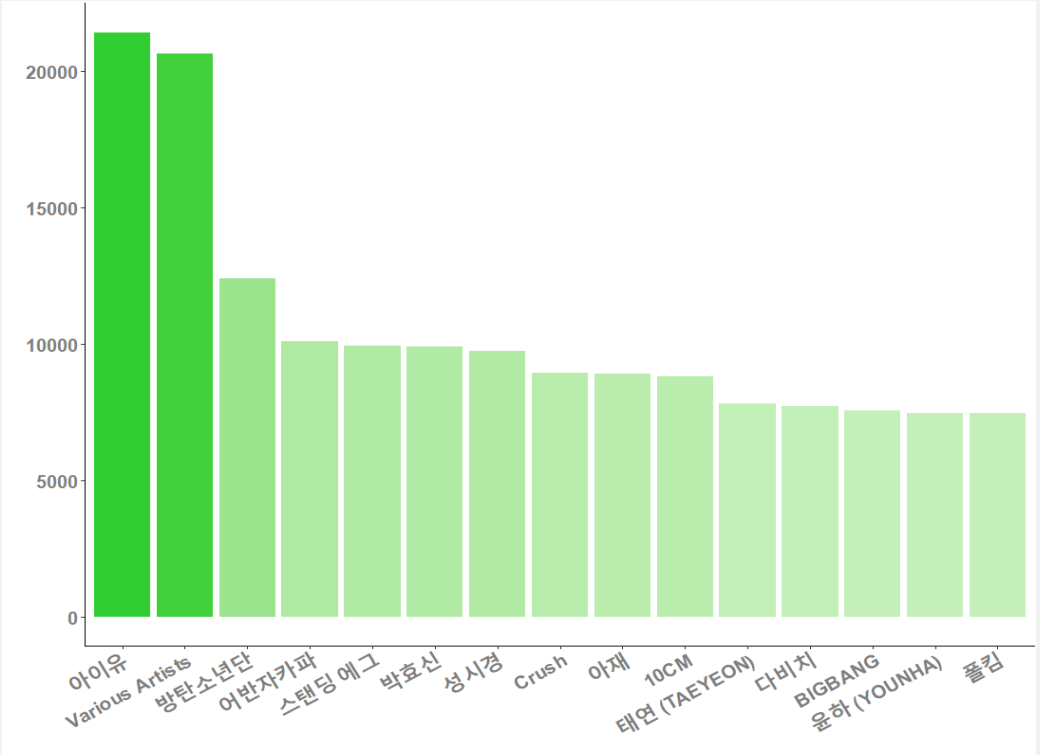
- 가장 많이 플레이리스트에 들어간 곡

곡	가수	횟수
밤편지	아이유	1219
안아줘	정준일	1184
비	폴킴	1109
그대와 나, 설레임	어쿠스틱 콜라보	904
야생화	박효신	891
어떻게 지내	Crush	881

잔잔/무난/대중적인 곡들이 사랑받는 중!

03. 데이터 탐색

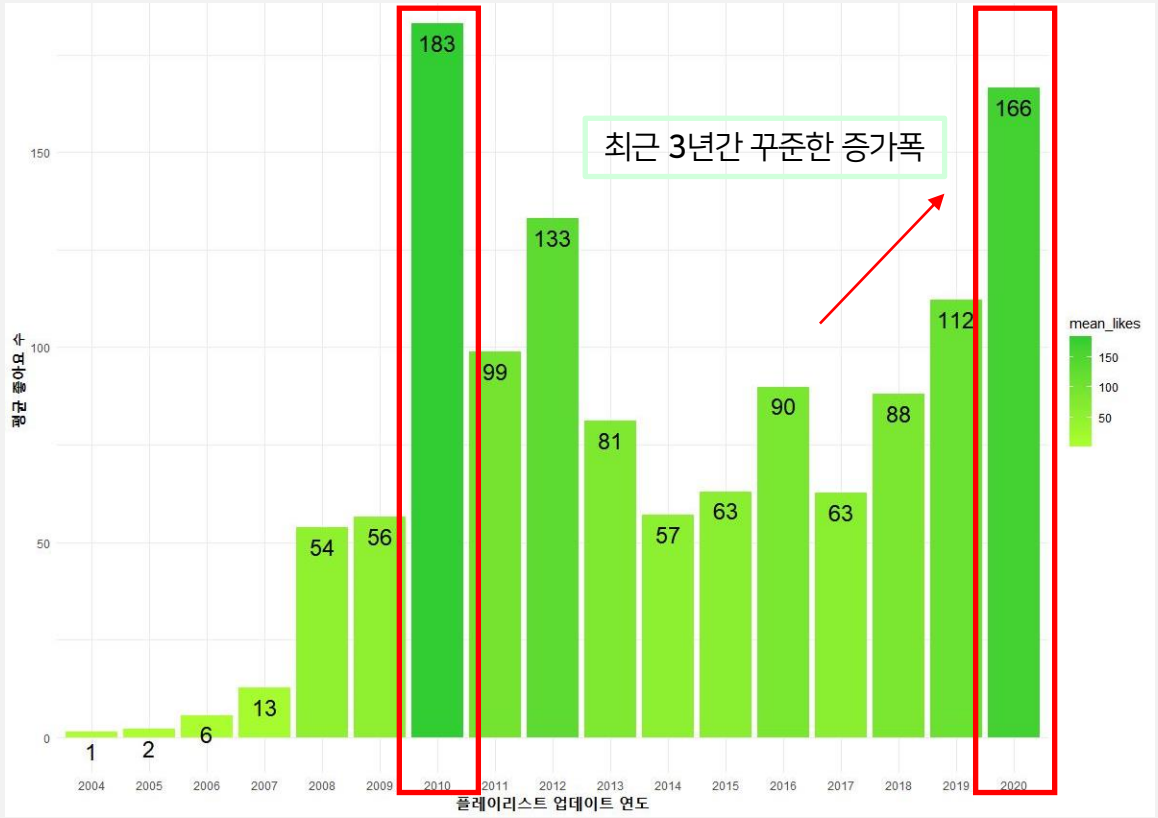
- 가장 많이 플레이리스트에 들어간 가수



Various Artists가 아이유의 자리를 위협 중!

03. 데이터 탐색

- 플레이리스트 업데이트 연도별 좋아요 개수의 평균



최근 업데이트 되었을수록 사랑받지만, 꾸준히 사랑받은 플레이리스트도 존재!

03. 데이터 탐색

- 플레이리스트 전체 태그 확인

playlist id	playlist name	태그
61281	여행같은 음악	['락']
10532	요즘 너 말야	['추억', '회상']
76951	편하게, 잔잔하게 들을 수 있는 곡.-	['까페', '잔잔한']
131982	퇴근 버스에서 편히 들으면서 하루를 마무리하기에 좋은 POP	['잔잔한', '버스', '퇴근버스', 'Pop', '풍경', '퇴근길']
100389	FAVORITE POPSONG!!!	['노래추천', '팝송추천', '팝송', '팝송모음']

하나의 플레이리스트 안에
다양한 태그들이 존재

가장 많이 붙은
태그는 어떤걸까?

03. 데이터 탐색

- 플레이리스트 전체 태그 확인

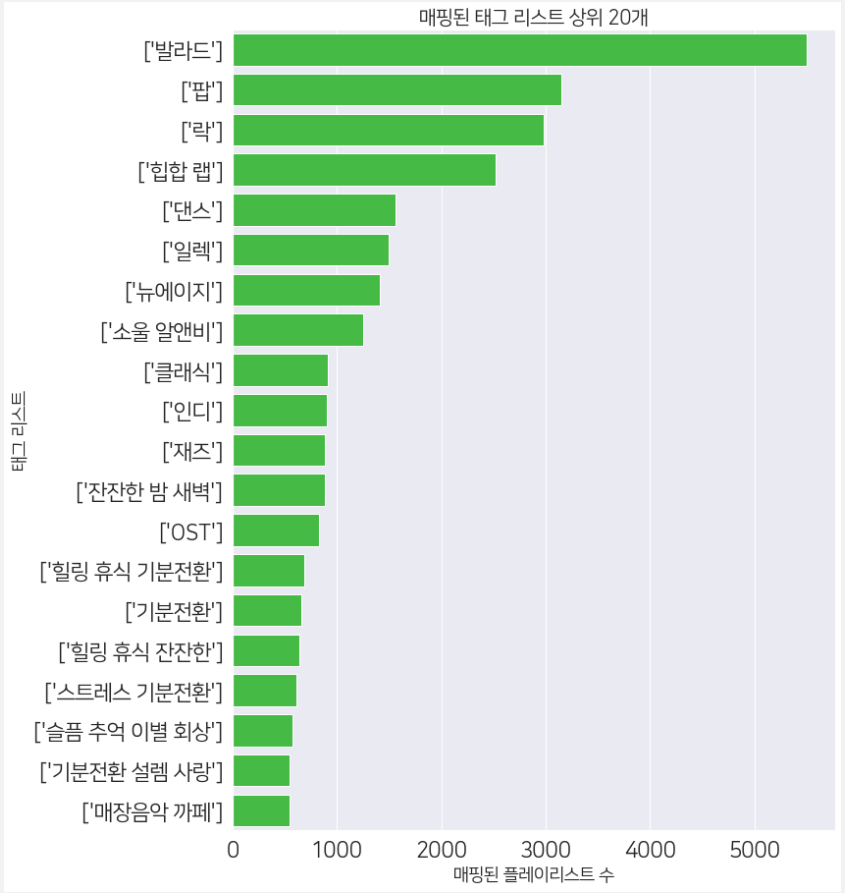
워드클라우드한 결과를 살펴보면

'기분전환', '감성', '휴식 ' 과 같은

태그들이 많은 모습이다

03. 데이터 탐색

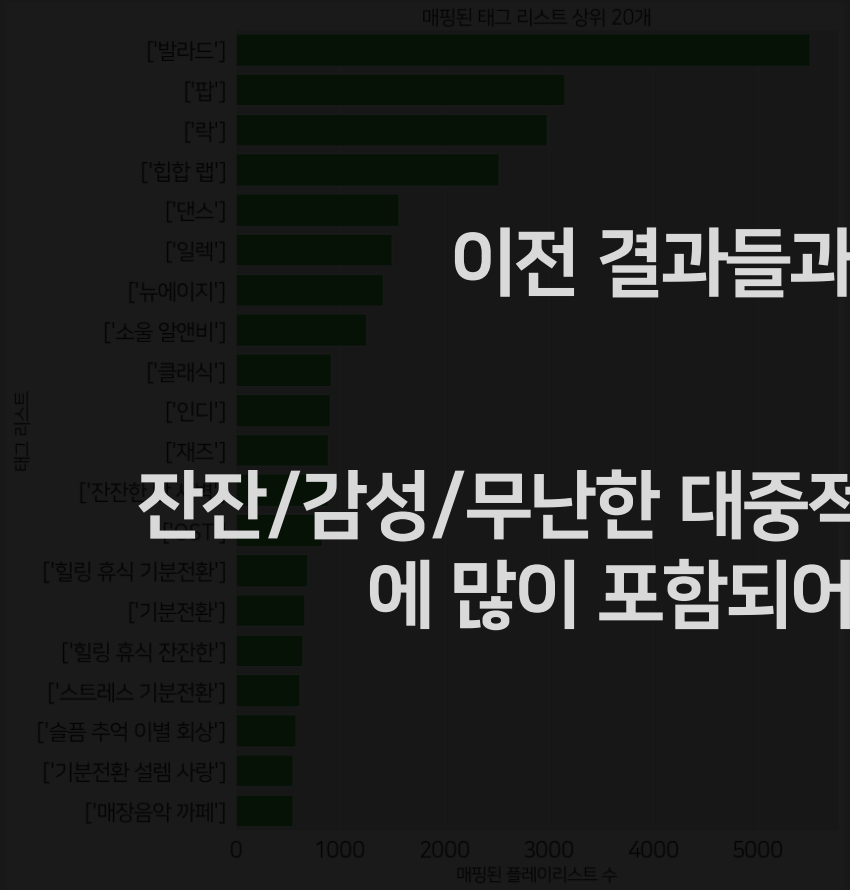
- 플레이리스트 전체 태그 확인



태그별로 플레이리스트에 몇 개가 매핑 되었는지 살펴보면 '발라드 ' 가 가장 많은 것을 확인할 수 있다

03. 데이터 탐색

- 플레이리스트 전체 태그 확인

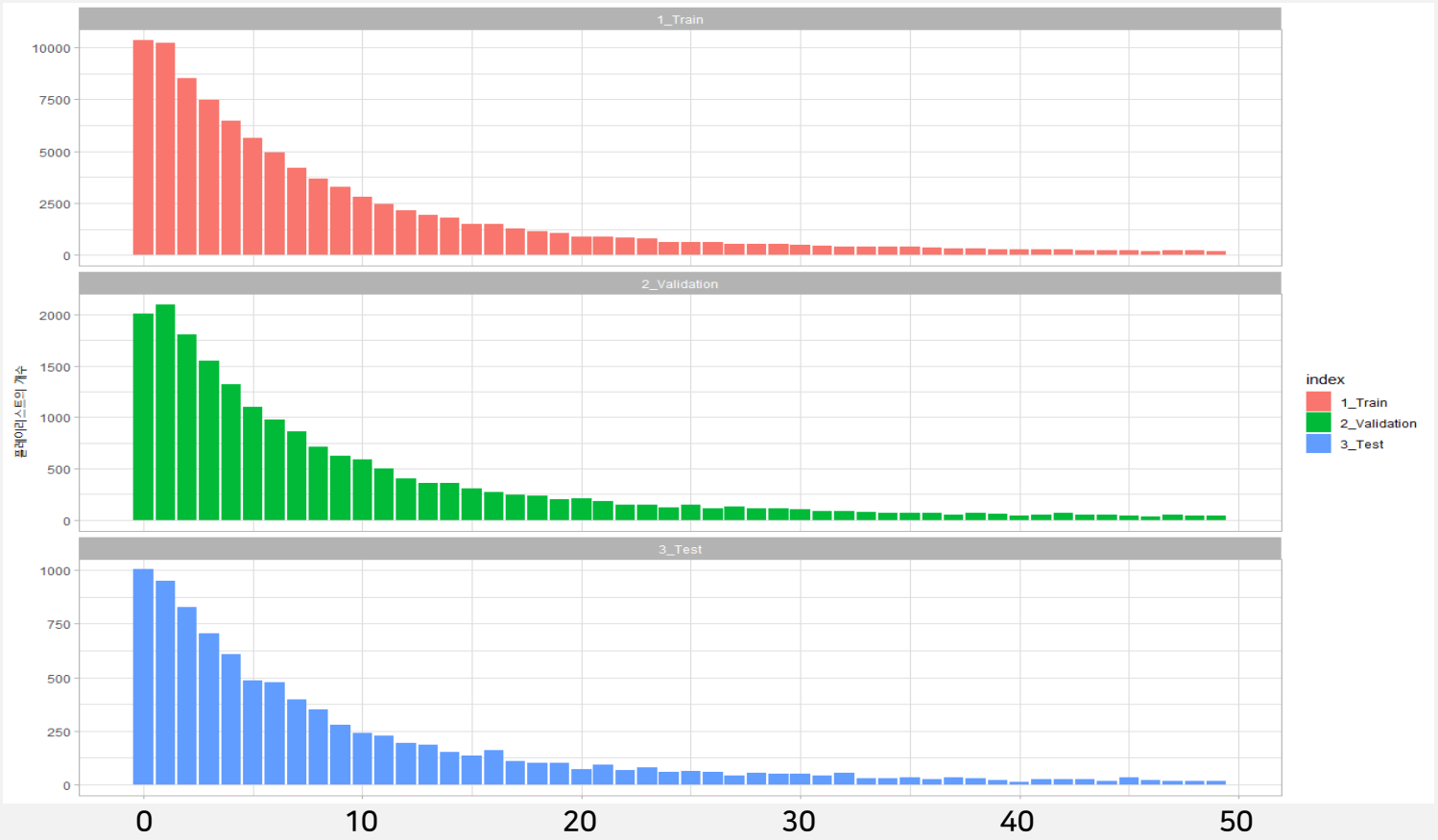


이전 결과들과 같이 이해하면,

잔잔/감성/무난한 대중적인 곡들이 플레이리스트들에 많이 포함되어 있음을 파악 가능!

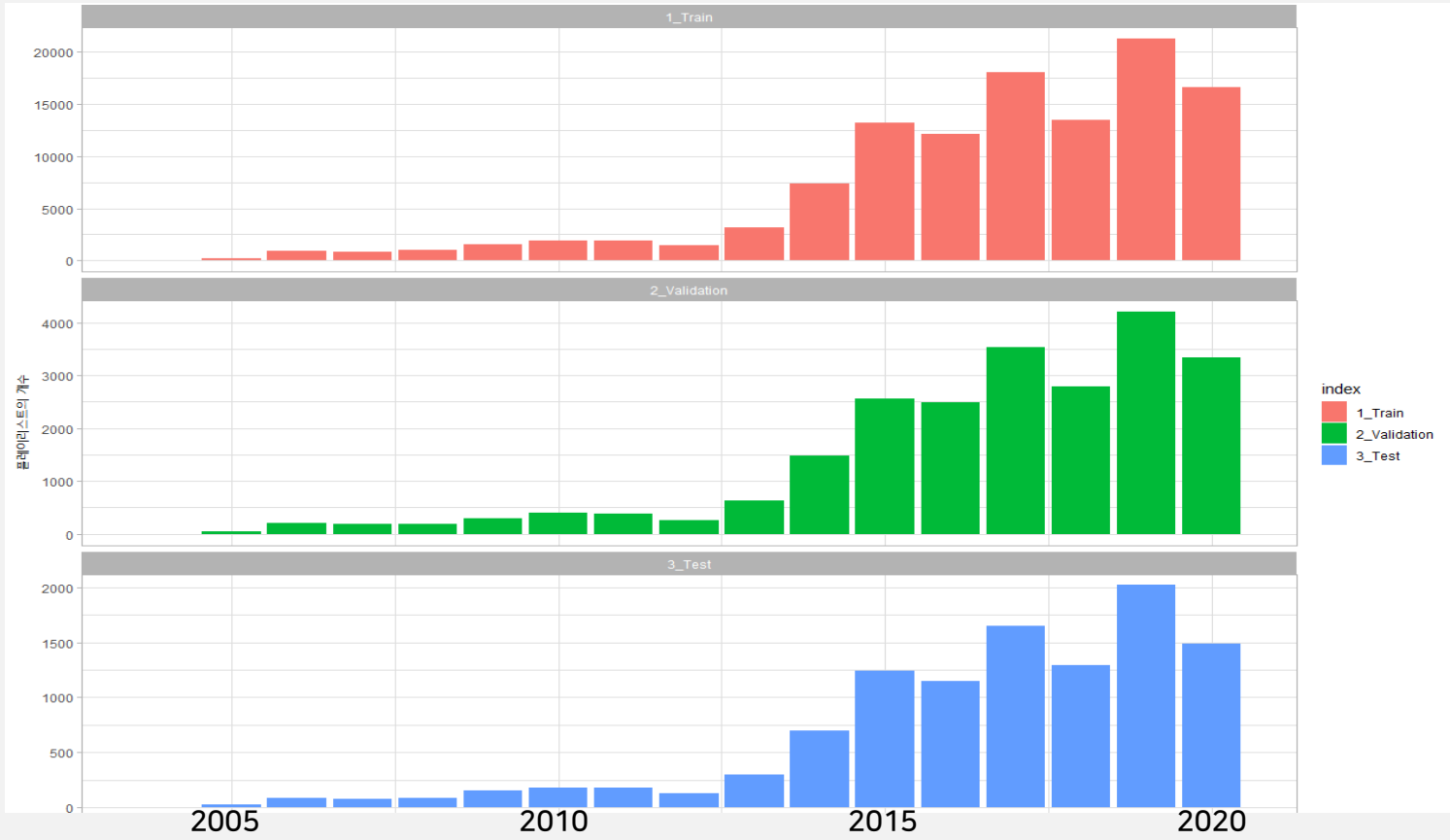
03. 데이터 탐색

- TVT 플레이리스트별 '좋아요 개수' 분포



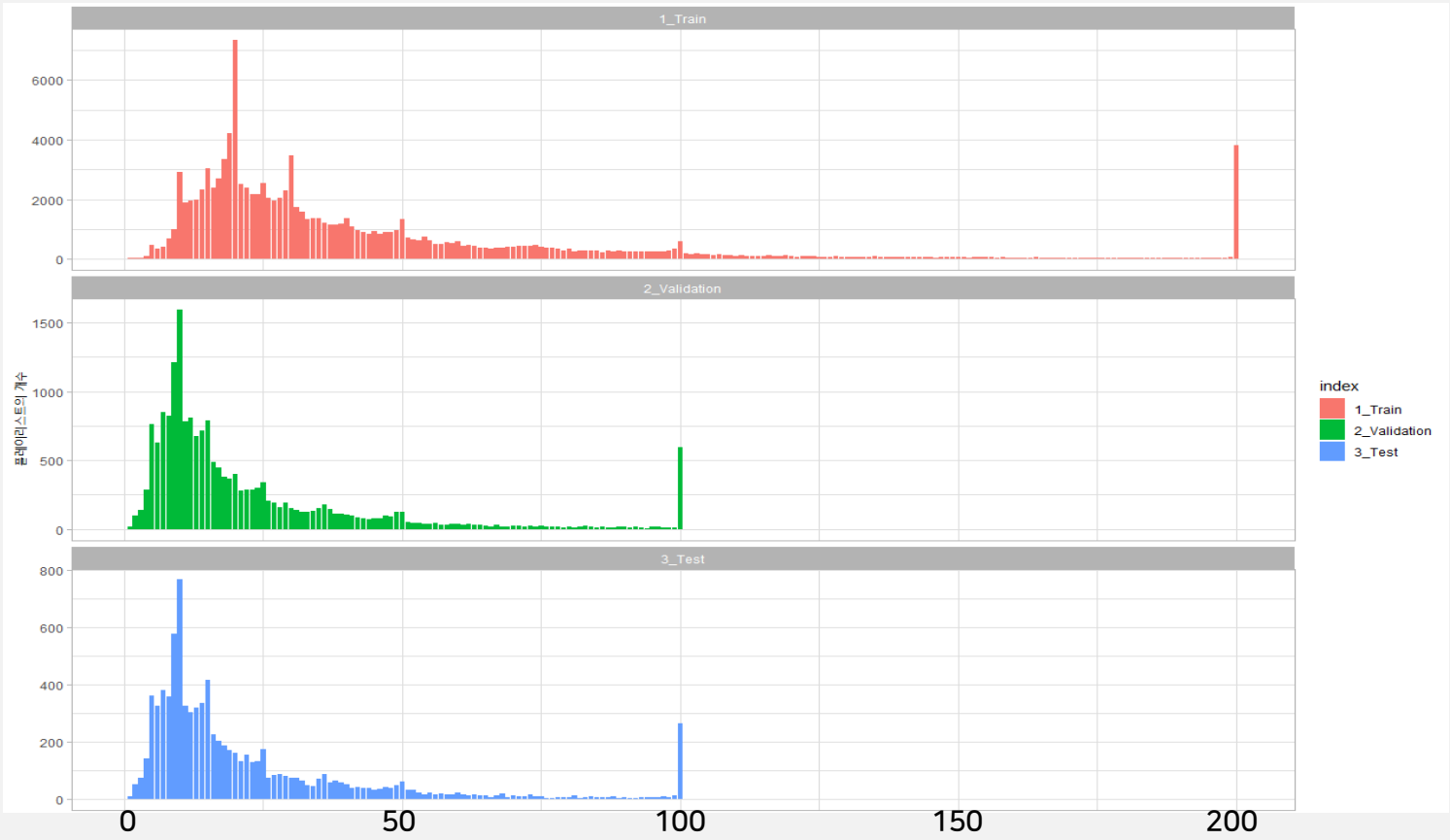
03. 데이터 탐색

- TVT 플레이리스트별 '업데이트 연도' 분포



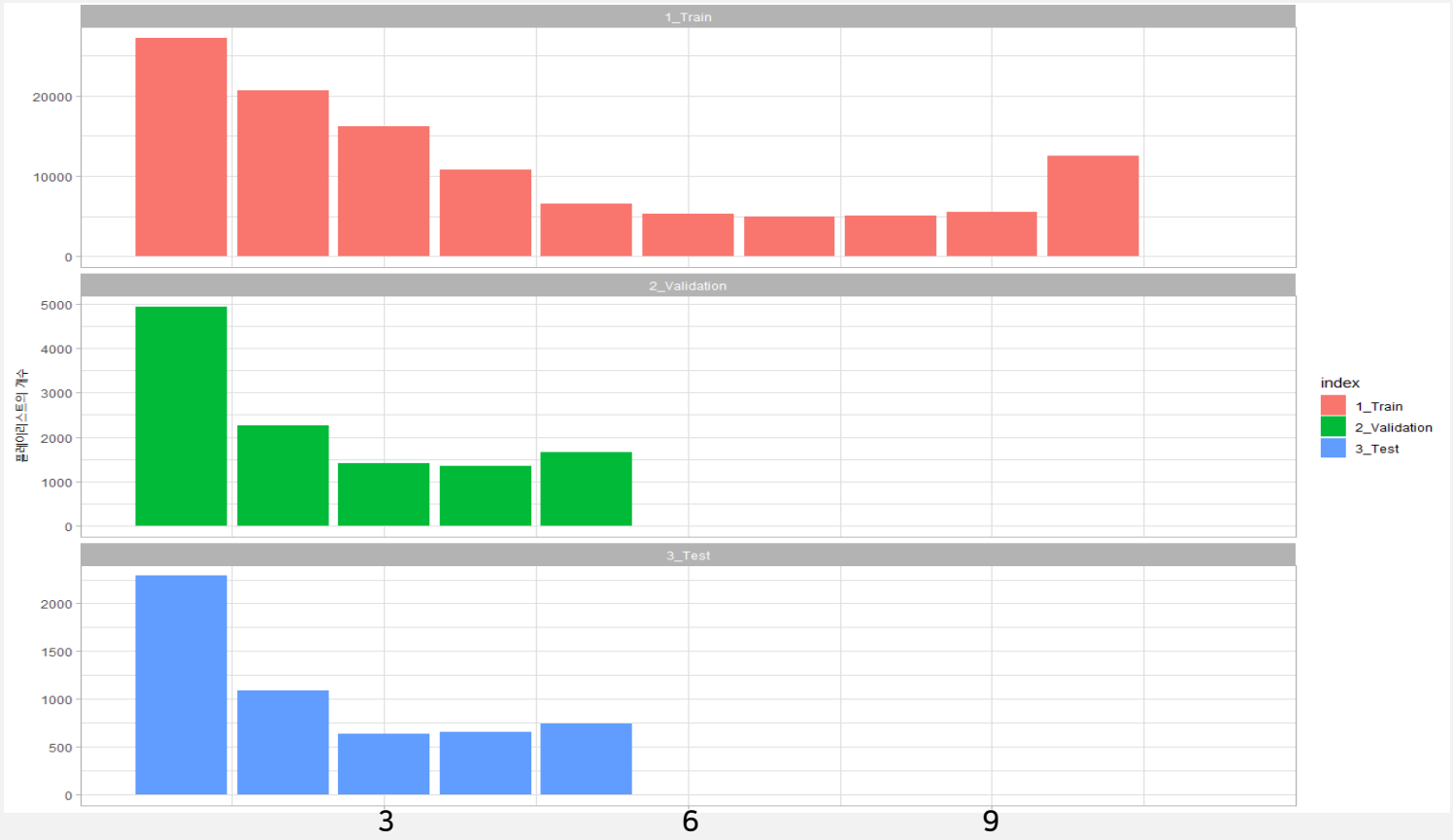
03. 데이터 탐색

- TVT 플레이리스트별 '곡 개수' 분포



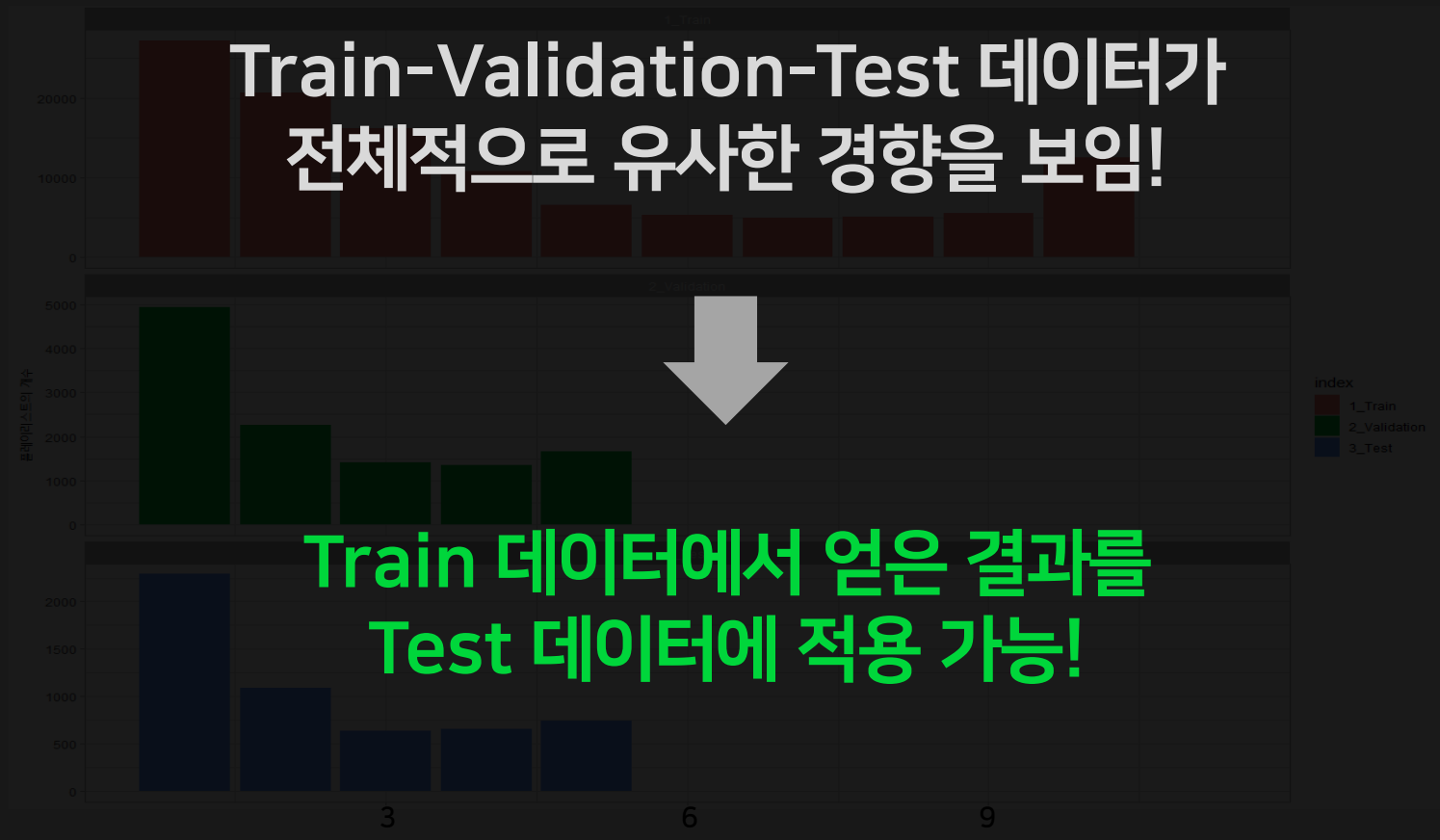
03. 데이터 탐색

- TVT 플레이리스트별 '태그 개수' 분포



03. 데이터 탐색

- TVT 플레이리스트별 '태그 개수' 분포



Train-Validation-Test 데이터가
전체적으로 유사한 경향을 보임!

Train 데이터에서 얻은 결과를
Test 데이터에 적용 가능!

03. 데이터 탐색

- 플레이리스트 태그에 대한 토픽모델링

태그	playlist id	playlist name	수록 곡 id	좋아요 개수	업데이트 일자
['락']	61281	여행같은 음악	[525514, 129701, 383374, 562083, 297861, 13954...	71	2013-12-19 18:36:19.000
['추억', '회상']	10532	요즘 너 말야	[432406, 675945, 497066, 120377, 389529, 24427...	1	2014-12-02 16:19:42.000
['까페', '잔잔한']	76951	편하게, 잔잔하게 들을 수 있는 곡.-	[83116, 276692, 166267, 186301, 354465, 256598...	17	2017-08-28 07:09:34.000
['잔잔한', '버스', '퇴근버스', 'Pop', '풍경', '퇴근길']	131982	퇴근 버스에서 편히 들으면서 하루를 마무리하기에 좋은 POP	[533534, 608114, 343608, 417140, 609009, 30217...	4	2019-10-25 23:40:42.000
['노래추천', '팝송추천', '팝송', '팝송모음']	100389	FAVORITE POPSONG!!!	[26008, 456354, 324105, 89871, 135272, 143548,...	17	2020-04-18 20:35:06.000

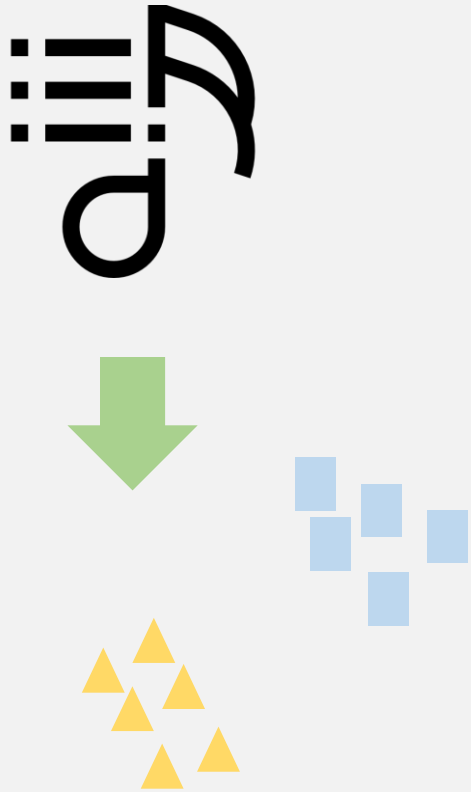
플레이리스트의 태그 관통하는 '토픽'들은 무엇이 있을까?

03. 데이터 탐색

- LDA (Latent Dirichlet Allocation)

✓ 비지도 학습 방법

✓ 토픽개수 k는 분석가가 결정
(혹은 perplexity로 결정)



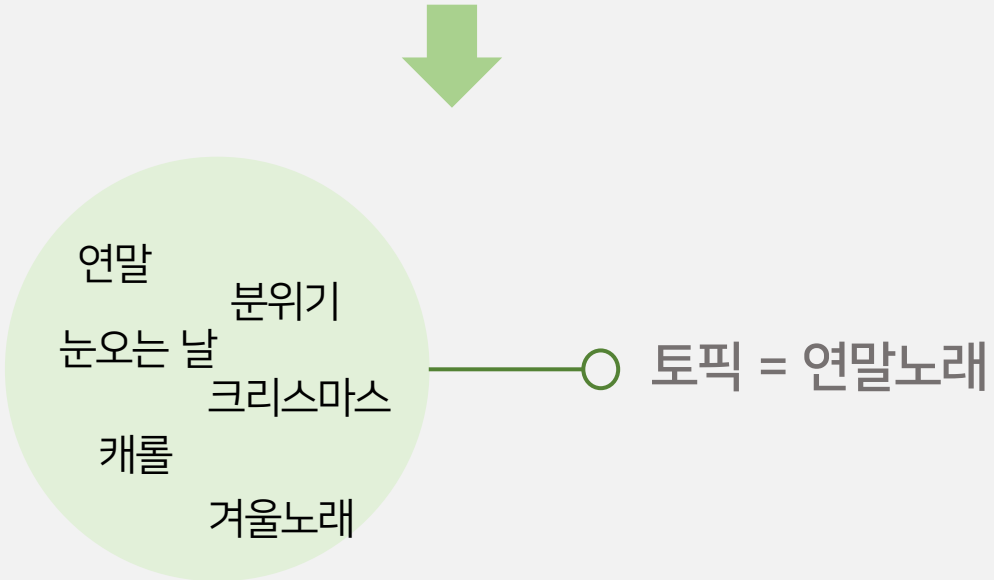
태그들을 잠재변수인 '토픽'의 실현으로 이해
각 토픽의 단어는 디리클레분포를 따름

03. 데이터 탐색

- LDA (Latent Dirichlet Allocation)

✓ 예시를 들어보자!

3	[연말, 눈오는날, 캐롤, 분위기, 따듯한, 크리스마스캐롤, 겨울노래, 크리스마스,...	147456	크리스마스 분위기에 흠뻑 취하고 싶을 때	[394031, 195524, 540149, 287984, 440773, 10033...	33	2019-12-05 15:15:18.000



03. 데이터 탐색

- LDA (Latent Dirichlet Allocation)

태그들을 전반적으로 관통하는 '토픽'들을 뽑아내보자

주어진 데이터로부터 주요한 특징들을 잘 잡아낸다면
새로운 정보가 더해졌을 때 매우 유용할 것!

Validation set, 또는 test set에서 태그에 대한 정보가 주어지지 않았을 때 **태그 예측**

Validation set, 또는 test set에서 수록곡에 대한 정보가 주어지지 않았을 때 **수록곡 예측**

03. 데이터 탐색

- LDA (Latent Dirichlet Allocation)
  - ✓ 컴퓨팅 파워는 고려해서…
  - ✓ 좋아요 200개 이상 + 2017년 부터의 플레이리스트

```
#좋아요를 200개 이상 받은 플레이리스트에서, 그 중 2017년 이후 발매된 플레이리스트만 다시 골라내기
indices=[]
for i in (list(many_likes.index)):
    our_date=tag_set['updt_date'].values[i][0:4]
    #print(i,our_date)
    if our_date=='2017' or our_date=='2018' or our_date=='2019' or our_date=='2020':
        indices.append(i)

#print(indices)
our_df=tag_set.iloc[indices]
our_df.head()
```

	tags	id	plylst_title	songs	like_cnt	updt_date
5	['운동', '드라이브', 'Pop', '트로피컬하우스', '힐링', '기분전환',...	69252	2017 Pop Trend	[418694, 222305, 96545, 135950, 304687, 457451...	435	2017-09-15 15:59:26.000
13	['힙합', '느낌있는', '밤', '새벽', 'RnB', '감각적인', '드라이브...	89809	트렌디하고 그루브한 힙합/알앤비 MUSIC	[525152, 38765, 66139, 696379, 397438, 144461,...	1112	2018-01-09 14:01:53.000
40	['위로', '힐링']	35178	아무것도 위로 되지 못할 때 위로가 되어준 음악	[512599, 308020, 124485, 296460, 218664, 31395...	1452	2019-06-26 19:49:20.000
61	['기분전환', '매장음악']	33019	취향저격 앨범아트	[97087, 682900, 490966, 703312, 64930, 417626,...	213	2017-03-29 19:51:15.000
102	['힙합', '라디오']	28452	악동서울DJ 힙합(NCT DREAM) 추천곡!	[241987, 67760, 205186, 32630, 335756, 547520,...	271	2019-08-21 21:55:43.000

```
our_df.shape[0]
```

```
6348
```

6348개의 플레이리스트 추출

03. 데이터 탐색

- LDA (Latent Dirichlet Allocation) 결과

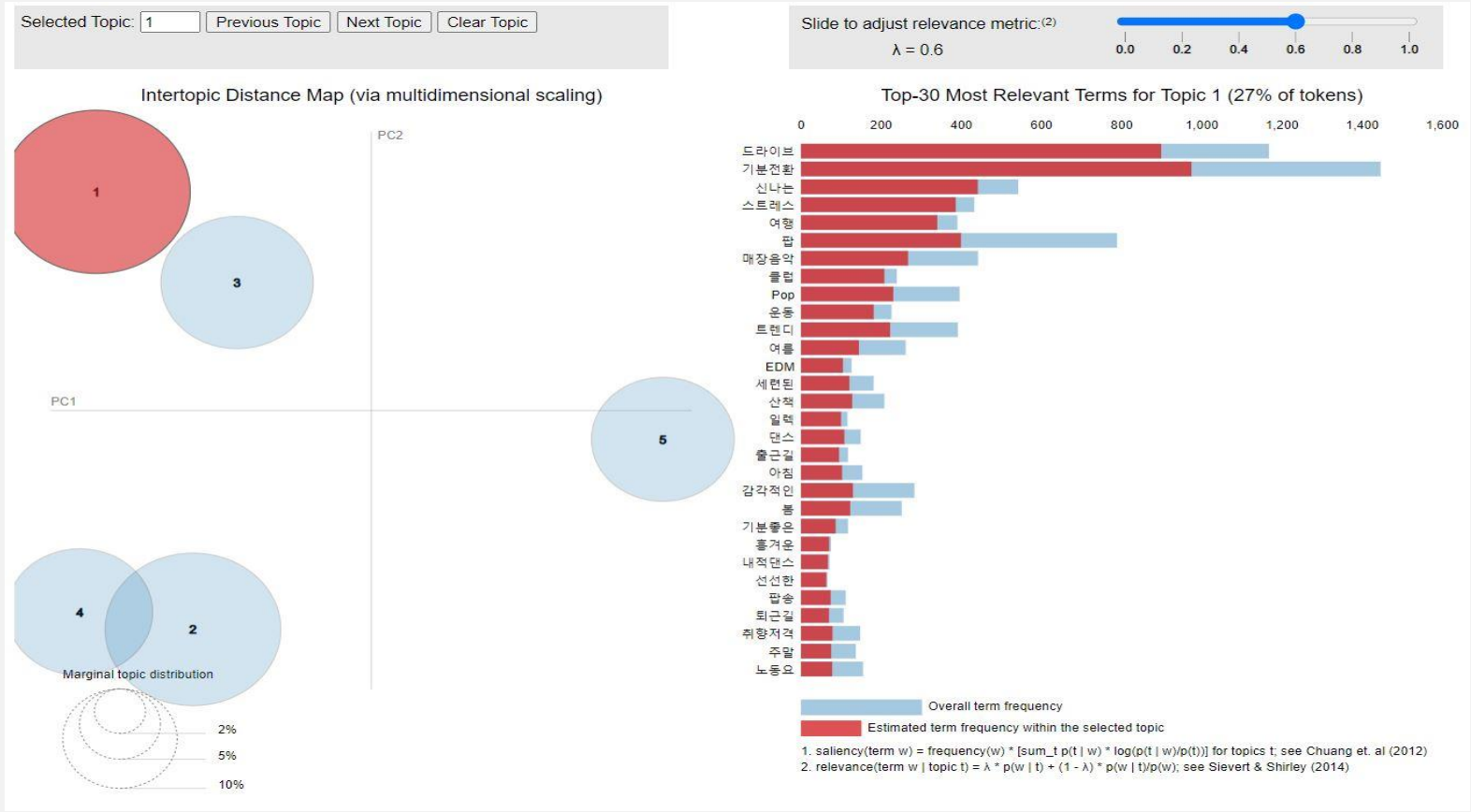
topic_num	word1	word2	word3	word4	word5	word6	word7	word8	word9	word10
0	기분전환	드라이브	팝	스트레스	신나는	여행	매장음악	트렌디	운동	사랑
1	휴식	힐링	기분전환	잔잔한	감성	카페	겨울	Pop	재즈	위로
2	밤	새벽	감성	잔잔한	인디	이별	가을	분위기	발라드	카페
3	피아노	뉴에이지	자장가	연주곡	OST	집중	힐링	휴식	공부	잔잔한
4	힙합	알앤비	기분전환	드라이브	추억	신나는	명곡	소울	회상	트렌디

다음과 같이 해당 단어가 토픽에 대해 지니는 가중치도 함께 산출

```
[(0,
  '0.072*"기분전환" + 0.063*"드라이브" + 0.040*"팝" + 0.036*"스트레스" + 0.034*"신나는" + '
  '0.026*"여행" + 0.024*"매장음악" + 0.024*"트렌디" + 0.020*"운동" + 0.020*"사랑"'),
 (1,
  '0.055*"휴식" + 0.054*"힐링" + 0.038*"기분전환" + 0.035*"잔잔한" + 0.031*"감성" + '
  '0.031*"카페" + 0.023*"겨울" + 0.019*"Pop" + 0.017*"재즈" + 0.016*"위로"'),
 (2,
  '0.068*"밤" + 0.067*"새벽" + 0.065*"감성" + 0.035*"잔잔한" + 0.033*"인디" + 0.028*"이별" '
  '+ 0.027*"가을" + 0.027*"분위기" + 0.025*"발라드" + 0.023*"카페"'),
 (3,
  '0.035*"피아노" + 0.029*"뉴에이지" + 0.024*"자장가" + 0.023*"연주곡" + 0.023*"OST" + '
  '0.023*"집중" + 0.022*"힐링" + 0.022*"휴식" + 0.018*"공부" + 0.017*"잔잔한"'),
 (4,
  '0.043*"힙합" + 0.037*"알앤비" + 0.036*"기분전환" + 0.031*"드라이브" + 0.030*"추억" + '
  '0.025*"신나는" + 0.018*"명곡" + 0.018*"소울" + 0.018*"회상" + 0.017*"트렌디"')]
```

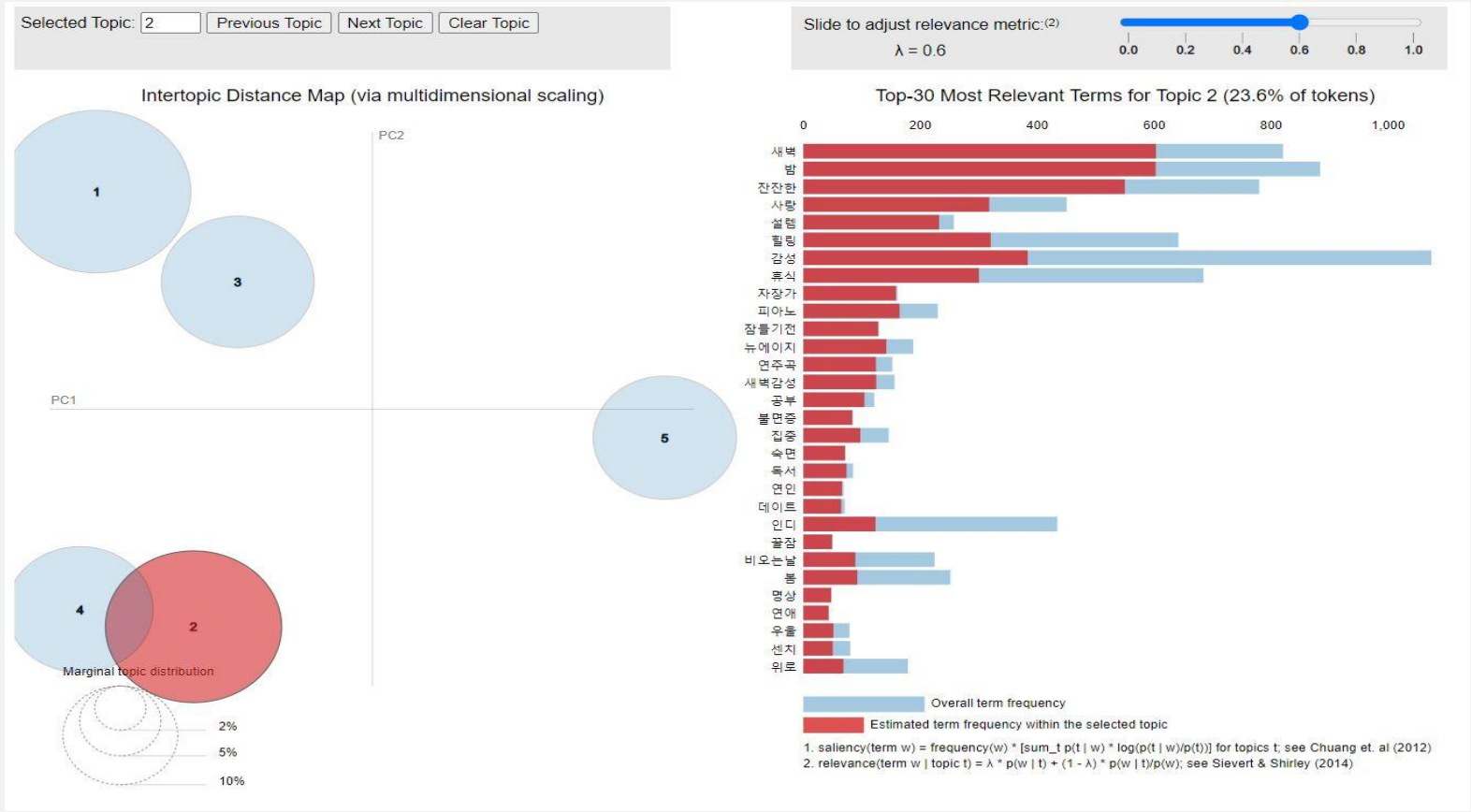
03. 데이터 탐색

- **토픽 1 : 에너제틱해서 기분전환하기 좋은 노래**



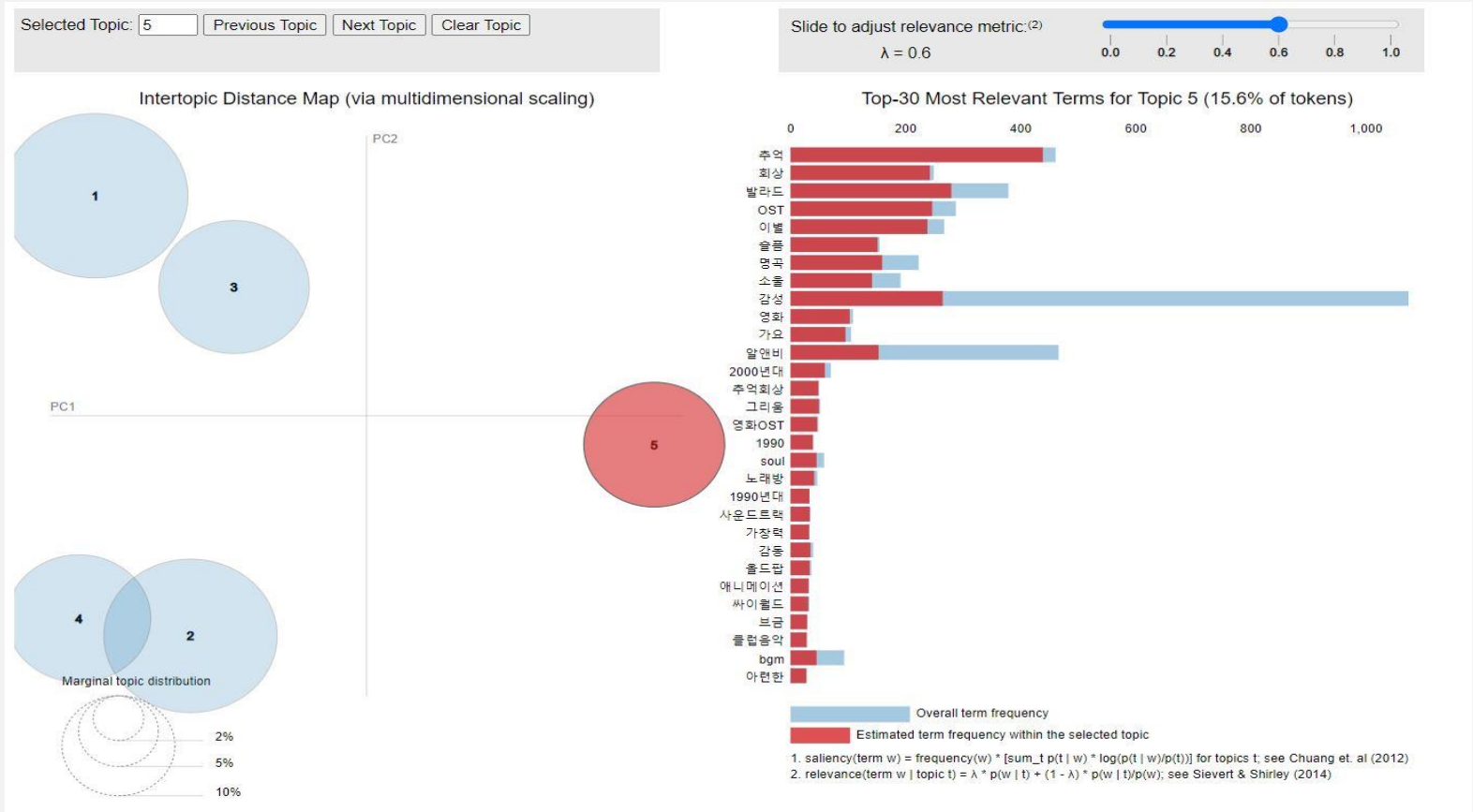
03. 데이터 탐색

- 토픽 2 : 잔잔한 감성의 노래



03. 데이터 탐색

- 토픽 5 : 추억의 노래



04. 3주차 예고

- Word2vec 임베딩
  - ✓ 문자를 컴퓨터가 이해할 수 있는 형식으로 변환

Local Representation	Continuous Representation
One-Hot encoding	Word2vec, NNLM
단어의 빈도수를 count	단어의 뉘앙스를 파악 가능
단어와 단어간의 관계를 이해할 수 없다.	

04. 3주차 예고

- Word2vec 임베딩

Word2vec이란 ?

두 단어의 유사한 정도를 이용해 주변 단어인지 아닌지를 예측하는 모델

Skip-gram

임베딩하려는 단어를 중심으로 다른 단어들이 그 단어 주변에 있는지 예측하는 모델

CBOW(Continuous Bag of Words)

주어진 단어에 대해 앞 뒤로 C/2개씩 총 C개의 단어를 사용하여 단어를 맞추는 모델

04. 3주차 예고

예시
"배가 고파서 나누미 떡볶이를 3인분을 먹었다."

CBOW 적용

"____ 고파서 나누미 떡볶이를 3인분을 먹었다."

빈 칸의 앞 뒤 문맥을 통해서
"배가" 가 들어가야함을 추론

Word2vec이란 ?

두 단어의 유사한 정도를 이용해 주변 단어인지 아닌지를 예측하는 모델

Skip-gram

CBOW(Continuous Bag of Words)

주어진 단어에 대해 앞 뒤로 C/2개씩
총 C개의 단어를 사용하여
단어를 맞추는 모델

04. 3주차 예고

- Word2vec 임베딩

예시
"배가 고파서 나누미 떡볶이를 3인분을 먹었다."

Word2vec이란 ?

두 단어의 유사한 정도를 이용해 주변 단어인지 아닌지를 예측하는 모델

Skip-gram 적용

Skip-gram

임베딩하려는 단어를 중심으로 다른 단어들이 그 단어 주변에 있는지 예측하는 모델

"배가 CBOW(Continuous Bag of Words) "

주어진 단어에 대해 앞 뒤로 C/2개씩
총 C개의 단어를 사용하여
"배가" 를 통해 나머지 문장을 예측

04. 3주차 예고

Word2vec 3주차 예고

- Skip-gram이 유의미할 수 있는 이유

태그	playlist id	playlist name	수록 곡 id	좋아요 개수	업데이트 일자
[]	118598		[373313, 151080, 275346, 696876, 165237, 52593...	1675	2019-05-27 14:14:33.000
[]	131447	앨리스테이블	[]	1	2014-07-16 15:24:24.000
[]	51464		[529437, 516103, 360067, 705713, 226062, 37089...	62	2008-06-21 23:26:22.00
['잔잔한']	101722		[75842, 26083, 244183, 684715, 500593, 508608,...	17	2015-12-17 14:06:05.000
['어머니', '힘들때', '아빠', '가족', '위로받고싶을때']	122127		[450275, 487671, 561031, 663944, 628672, 59121...	10	2020-04-16 21:35:44.000

일부의 정보를 가지고 비어있는 내용을 예측하는 맥락!

04. 3주차 예고

Word2vec 3주차 예고

- 태그 유의어 사전 만들기

```
from gensim.models import Word2Vec
model = Word2Vec(tags, size=20, window=2, min_count=10, workers=4, sg=1)
```

```
print(model.wv.most_similar('기분전환',topn=50))
```

```
[('카페에서듣기좋은', 0.5211527347564697), ('집에가', 0.516815721988678), ('신비', 0.5163450837135315), ('클래식명곡', 0.47231876850128174), ('Miguel', 0.4710850119590759), ('요리노래', 0.46989667415618896), ('데니스브레인', 0.46919432282447815), ('1414', 0.468281090259552), ('거리노래방', 0.46370169520378113), ('화지', 0.46369242668151855), ('멋지다', 0.46191370487213135), ('국악탑백', 0.46182698011398315), ('여름이다가온다', 0.4481390118598938), ('디제이음악', 0.44322752952575684), ('아당', 0.4431780278682709), ('JAPAN', 0.4425484538078308), ('나야나', 0.44213756918907166), ('그대로', 0.43713051080703735), ('힙플페', 0.43363332748413086), ('이별위로', 0.43114960193634033), ('일렉트로재즈', 0.4310801029205322), ('버닝썬', 0.428158700466156), ('유니티', 0.42748391628265538), ('바람에날려', 0.4253860116004944), ('TIFFANY', 0.4253017008304596), ('재즈바', 0.42395228147506714), ('악기', 0.4236276149749756), ('나서연', 0.4232034981250763), ('루와', 0.42133718729019165), ('찾아라new', 0.41873809695243835), ('Elephant', 0.4186609387397766), ('도시적', 0.41775232553482056), ('한국인이_사랑하는_팝송', 0.4171823263168335), ('아티쇼', 0.4168916344642639), ('최고의작곡가바흐인기클래식', 0.4158231317996979), ('베를린필', 0.4156760573387146), ('CALUMSCOTT', 0.41300103068351746), ('gershwin', 0.41294795274734497), ('새벽4시', 0.4116142690181732), ('찌릿찌릿', 0.4115907847881317), ('고은', 0.4112975001335144), ('기쁜', 0.4092404544353485), ('굿나잇송', 0.40901318192481995), ('찬영', 0.40827134251594543), ('고즈넉한', 0.4081392288208008), ('Malice', 0.40755945444107056), ('퍼블릭에너미', 0.4054603576660156), ('2010년초반', 0.4053986966609955), ('노출', 0.4038392901420593), ('평화를', 0.4035341739654541)]
```

플레이리스트의 태그들을 Word2Vec으로

유사한 태그들의 '코사인 유사도' 산출

04. 3주차 예고

- 태그 유의어 사전 만들기

✓ 상위 50개 태그 확인

```
m_tag=most_tag_cnt[0:50]
top_50_tag=m_tag[['tags']].values
top_50_tag
```

✓ 코사인 유사도가 0.6 이상인 태그 추출

```
popular_tag=[]
for tag in top_50_tag:
    top_50_similiar = []
    temp = model.wv.most_similar(tag, topn=50)
    for lst in temp:
        if lst[1] > 0.6:
            top_50_similiar.append(lst[0])
    popular_tag.append([tag, top_50_similiar])

#      popular_tag.append([tag, model.wv.most_similar(tag, topn=50)])
df = pd.DataFrame(data=popular_tag,
                  columns=['tag', 'similarity'])
```

04. 3주차 예고

Word2vec 3주차 예고

- 태그 유의어 사전 만들기

	tag	similiarity
0	[기분전환]	[한강, 아이돌그룹, 스밍, 비오는날엔, 얼터너티브록, 귀성길, 멜론스포츠, 신나는음악]
1	[감성]	[무한도전, 살랑살랑]
2	[휴식]	[달달한노래, 0살, 랩발라드, 향수, 공항, 팬시차일드]
3	[발라드]	[추억여행, 신비, 팝송을, 송라이터, 어린이클래식, 청량함, 남자친구, 팝송명곡]
4	[잔잔한]	[잔잔함, 추운날, 해외차트, 마블, 해외음악, 만화주제가]
5	[드라이브]	[봄노래, 멜랑꼴리, 나이, 모음집, 집에가는길, 길거리, 초여름]
6	[힐링]	[작곡가, 소나기, 8월, 고요, 집에서, 스타일리시, 나만알고싶은노래, 빅밴드]
7	[사랑]	[러블리즈, 방콕, 우주소녀, COOL, 추운, 후유증, 헬스음악, 편집샵, 평온]
8	[새벽]	[창모, 프로듀싱, 5월, 뜨거운, 텐션업, 봄소풍]
9	[밤]	[크로스핏, 헬스장, 레이디가가, 인생노래, 불토, 달리기, 밤에, 클래식음악]
10	[카페]	[유니크한, 슬로우잼, 새로운시작, 박효신]
11	[추억]	[비오는_날, 질리지않는, 위너, 디저트, 밤산책, Labelpick]
12	[팝]	[여유로운, 프로듀서, 클럽, EDM모음, 한동윤, 늦여름, 레이디가가, 에프엑스]

04. 3주차 예고

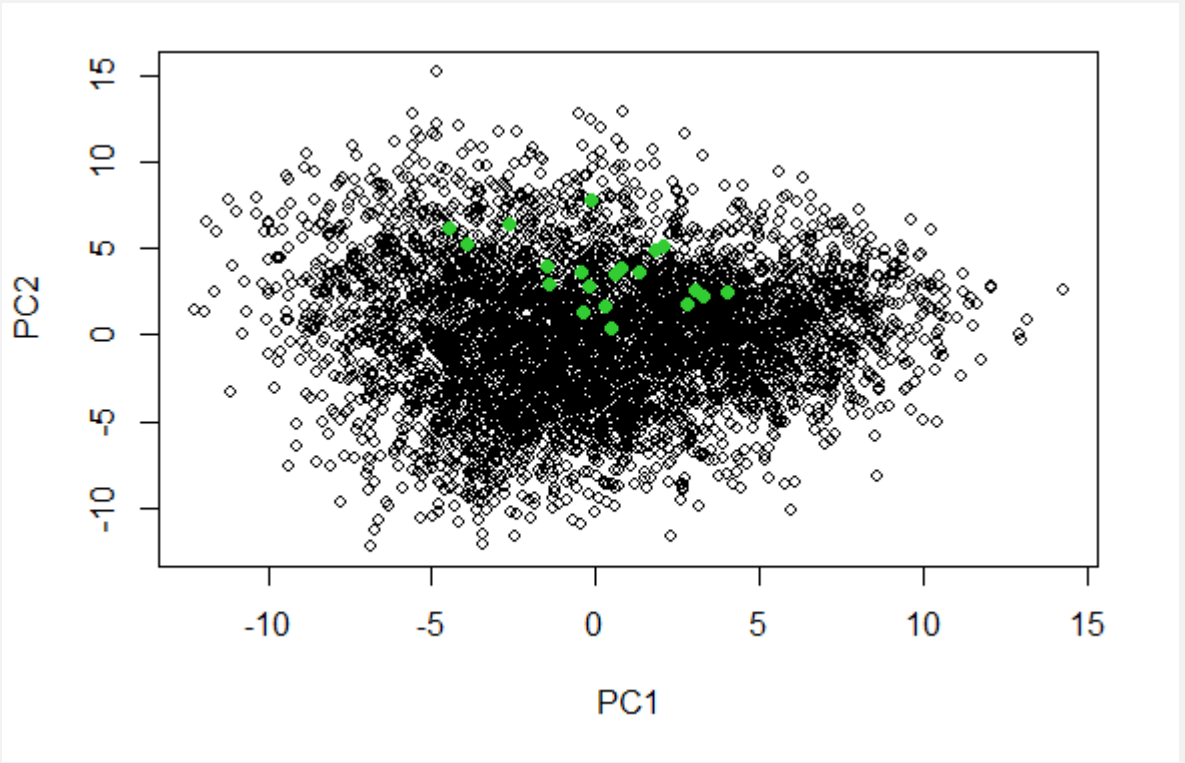
Word2vec 3주차 예고

- 태그 유의어 사전 예시

	tag	similarity
0	기분전환	한강, 신나는 음악, 얼터너티브 록, 아이돌그룹, 신나는…
1	감성	살랑살랑…
2	휴식	달달한 노래, 향수…
3	발라드	추억여행, 남자친구, LOVE, 잔잔한 팝송,…
4	잔잔한	잔잔함, 추운날…
5	드라이브	집에 가는길, 길거리, 초여름, 멜랑꼴리…
6	힐링	고요, 집에서, 나만알고싶은 노래,
7	사랑	러블리즈,
8	새벽	새벽녘, 고요한, 창모…
9	밤	밤에, 클래식 음악…
…	…	…

04. 3주차 예고

- 플레이리스트 태그의 word2vec 결과

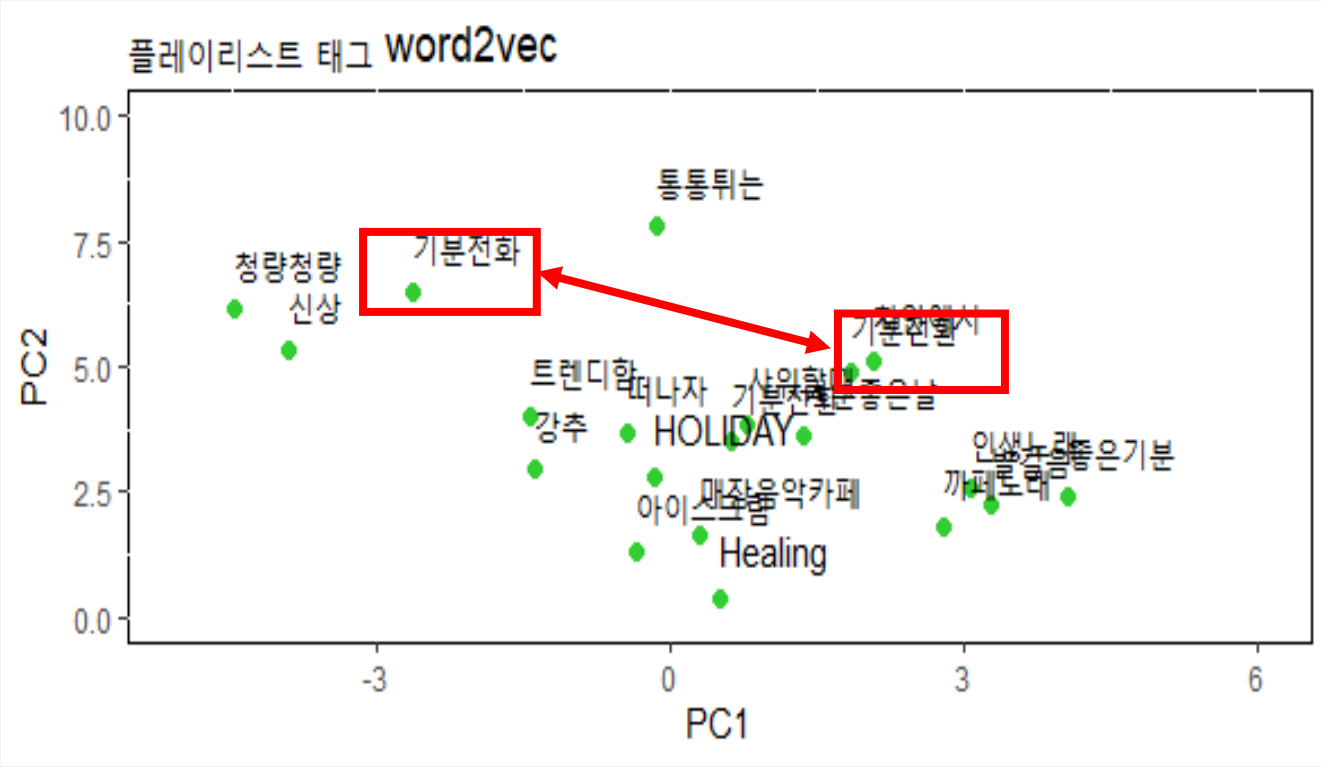


'기분전환'의 유사어를 이차원에서 시각화한 결과

04. 3주차 예고

Word2vec 3주차 예고

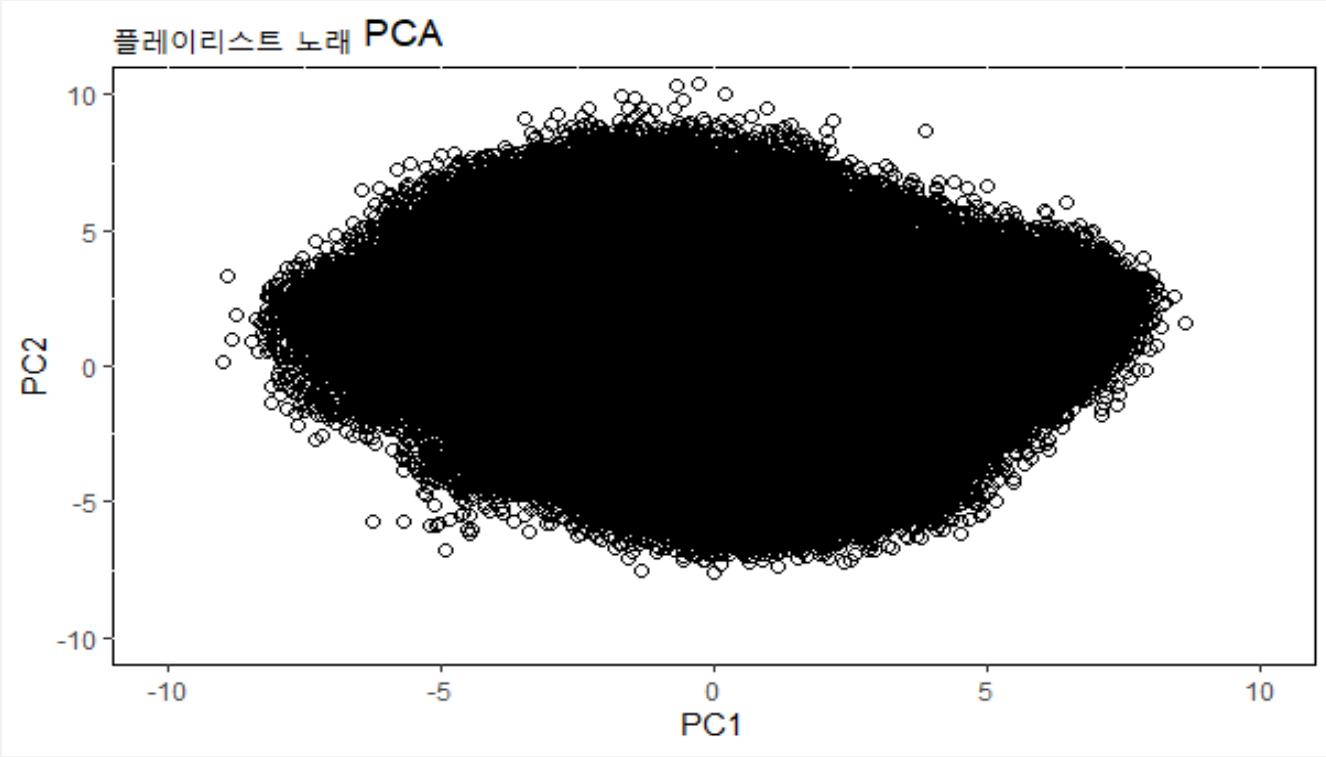
- 플레이리스트 태그 내 유사어 PCA



100차원의 코사인 유사도를 계산했지만, 이차원 상에서 거리를 확인해봄

04. 3주차 예고

- 플레이리스트 노래 word2vec 결과 PCA

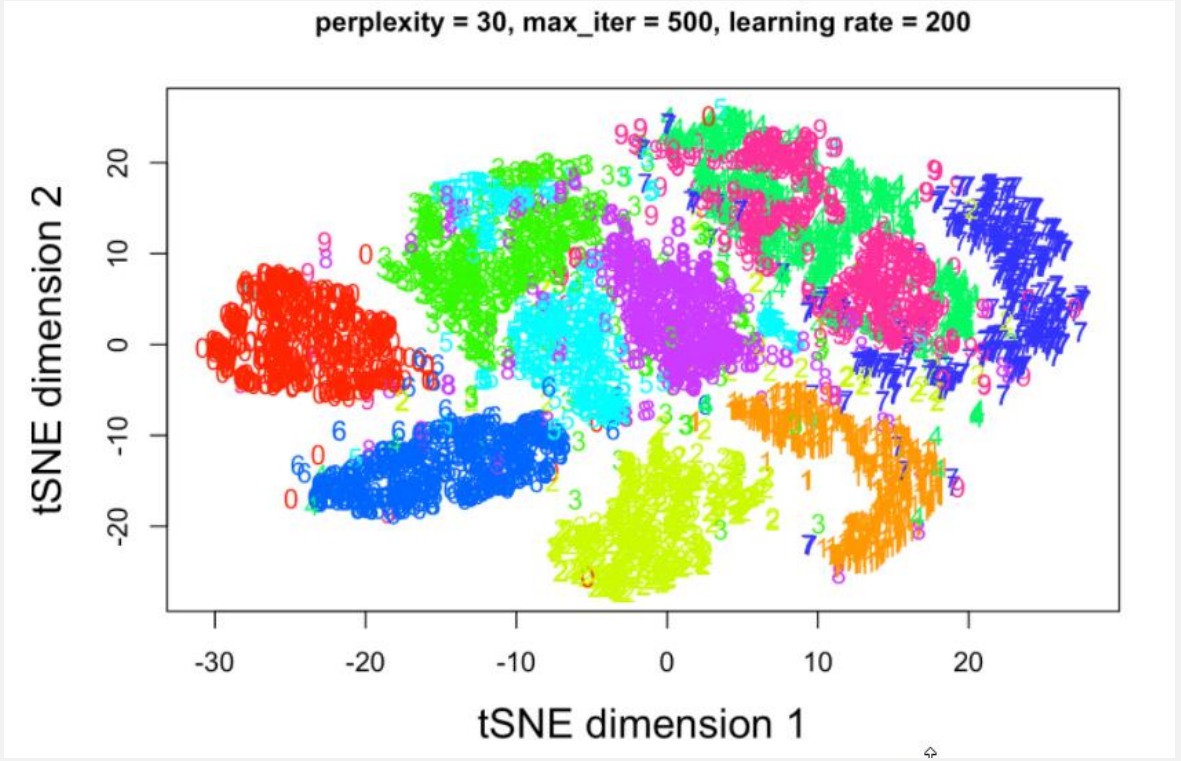


고차원상에서 유의미한 다양체 구조를 PCA로 파악하는 것은 매우 어려움…!

04. 3주차 예고

Word2vec 3주차 예고

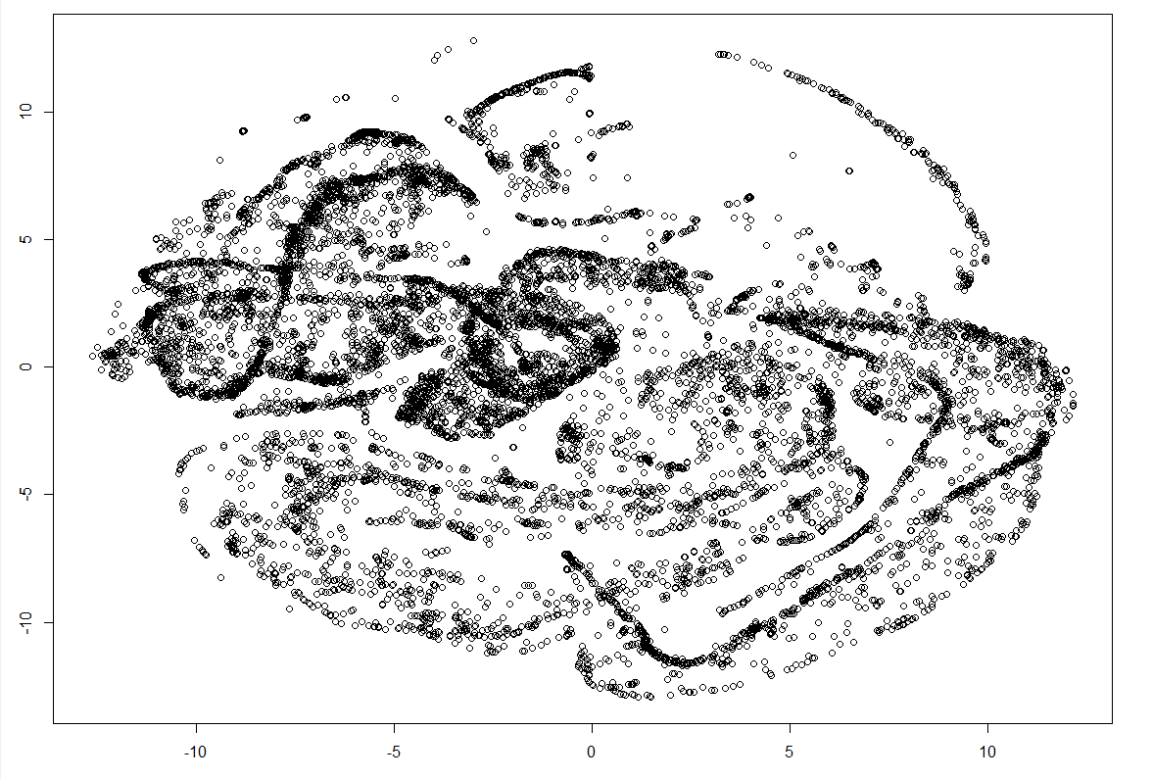
- tSNE (t-distributed Stochastic Neighbor Embedding)



고차원상에서 유의미한 다양체 구조를 이차원 상에서도 유사하도록 보존해주는 차원 축소 방법!

04. 3주차 예고

- 플레이리스트 곡 tsne 시각화, 하지만…



명확한 군집형태가 드러나지 않음!

04. 3주차 예고

Word2vec 3주차 예고

- tsne 시각화가 잘 작동하지 않는 이유

태그	playlist id
['락']	61281
['추억', '회상']	10532
['까페', '잔잔한']	76951
['잔잔한', '버스', '퇴근버스', 'Pop', '풍경', '퇴근길']	131982
['노래추천', '팝송추천', '팝송', '팝송모음']	100389

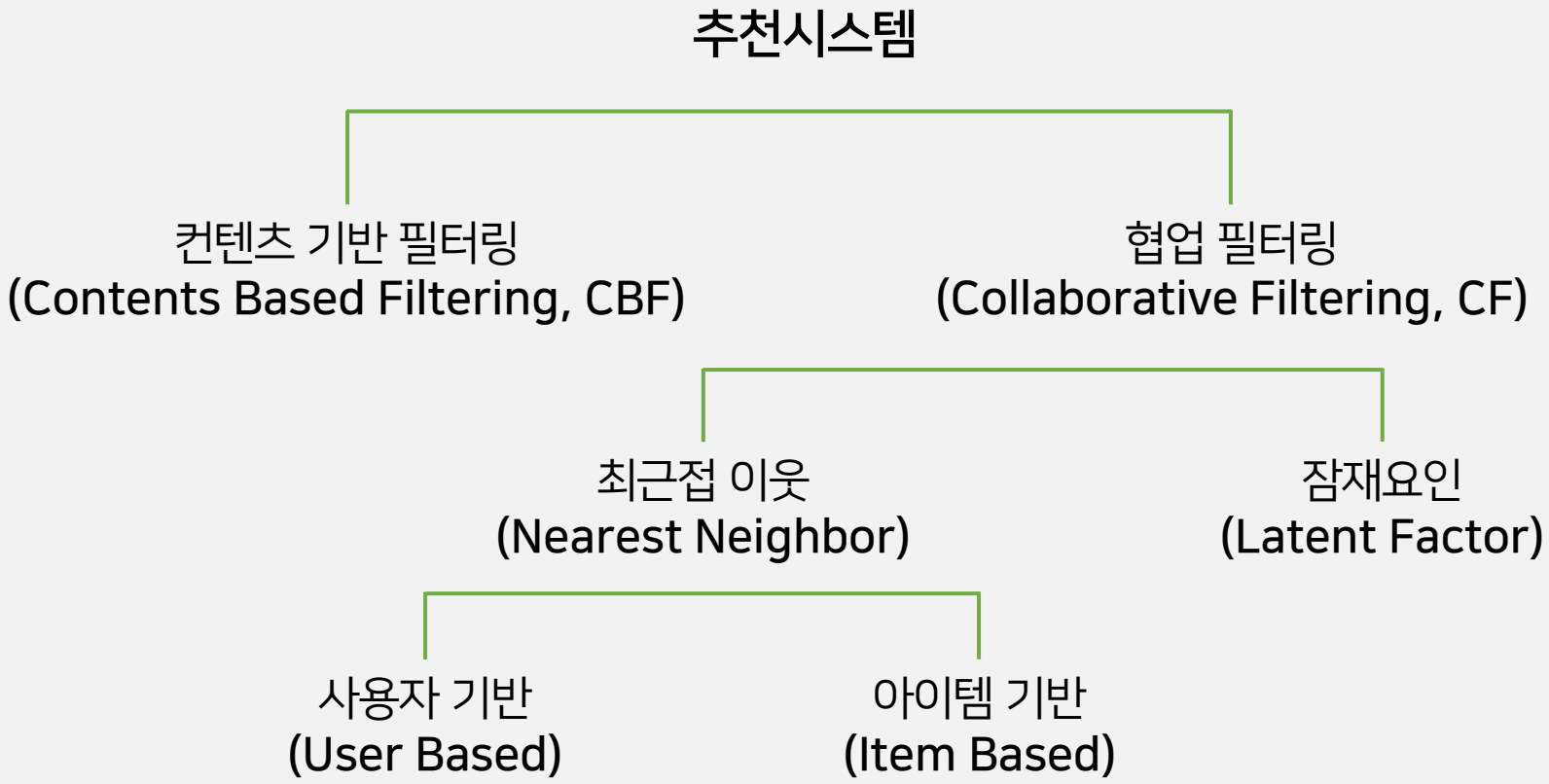
class	article
film	After over monthlong delay movie Time Hunt w...
film	years after Sewol ferry sank dark bottom Apr...
film	With ever dropping number moviegoers three m...
film	number weekend moviegoers continued downside...
film	people continue isolate themselves inside ho...
...	...
politics	reaffirmation came during National Police Ag...
politics	would like have opportunity meet soon share ...
politics	Election campaign pledges actual state affai...
politics	which primary parties that have contributed ...
politics	rival parties have agreed hold their vote ap...

이런 형태에는 잘 작동

한 플레이리스트가 가진 태그가 적기 때문에 잘 묶지 못한다!

04. 3주차 예고

- 다음주에는…?



04. 3주차 예고

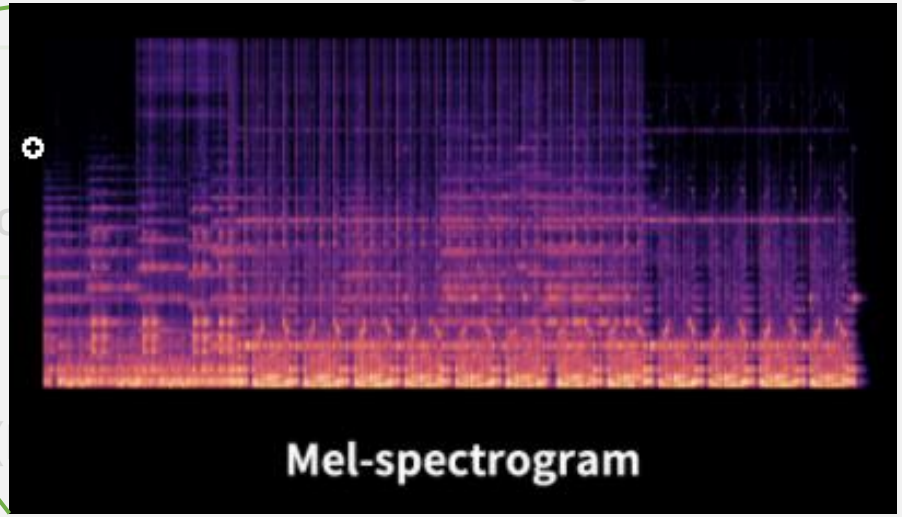
추천시스템

컨텐츠 기반 필터링
(Contents Based Filtering, CBF)

협업 필터링
(Collaborative Filtering, CF)

최근접 이웃
(Nearest Neighbo

사용자 기반
(User Based)



240 GB 데이터……

04. 3주차 예고

20:00 셔플재생

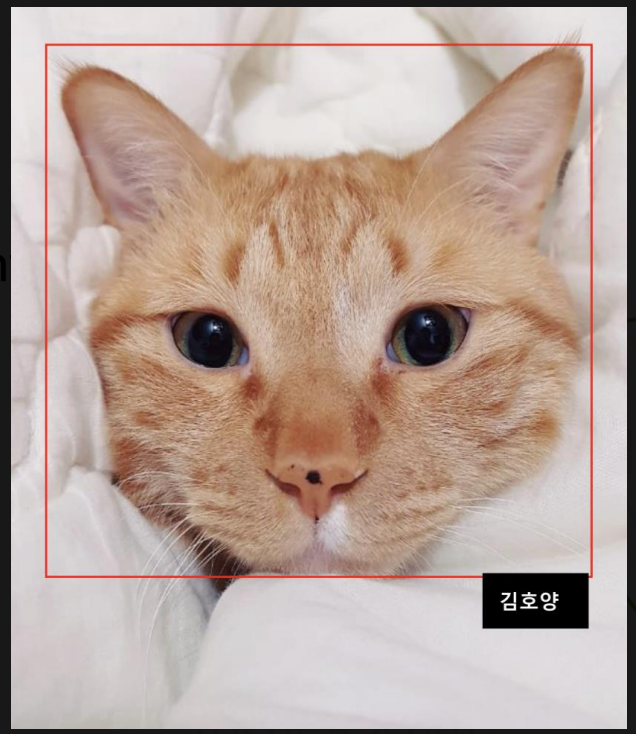
01. 주제 선정

02. 데이터 확인

03. 데이터 탐색

3주차 예고

(Con



딥러닝팀 본체 호양이

04. 3주차 예고