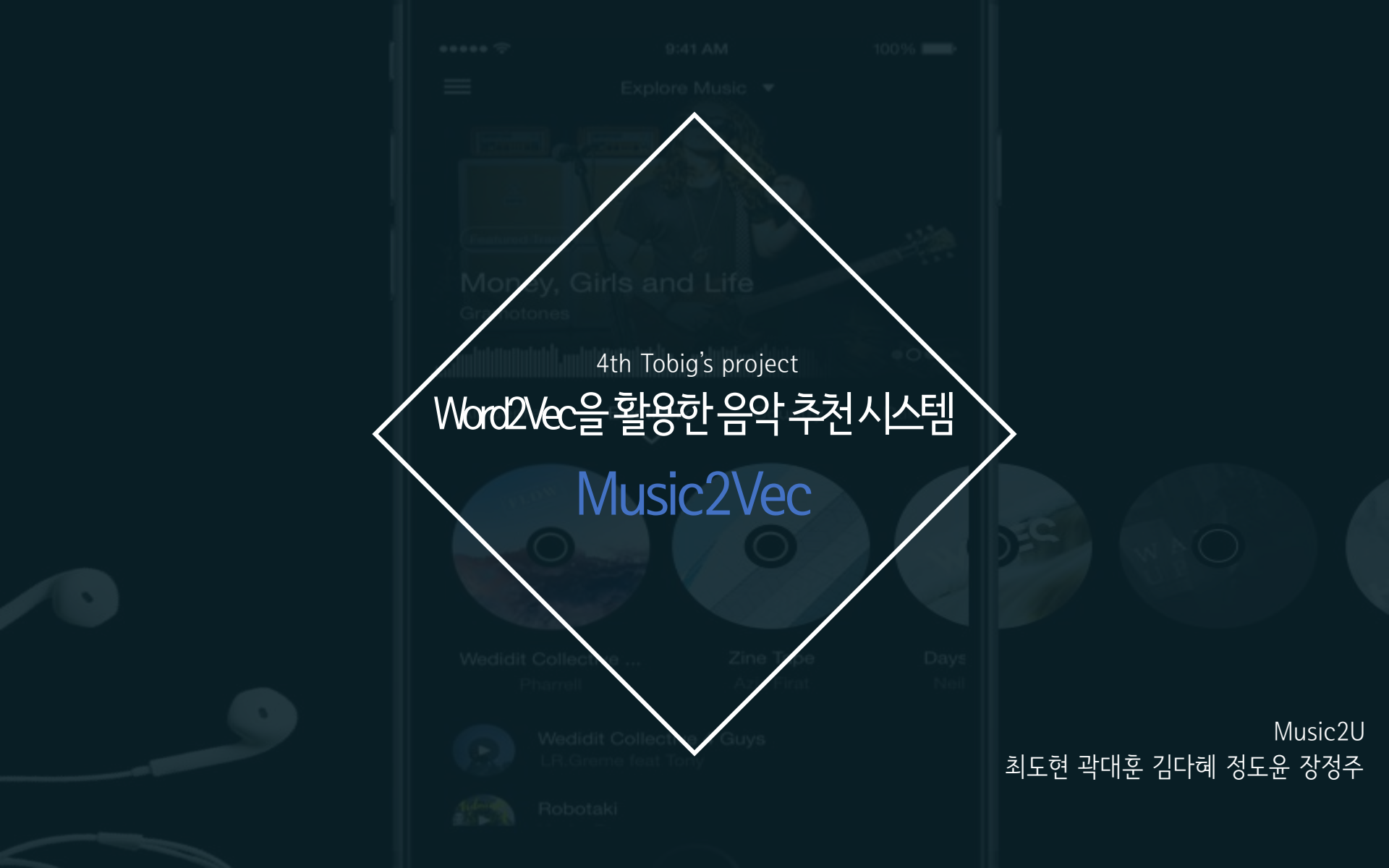


4th Tobig's project

Word2Vec을 활용한 음악 추천 시스템

Music2Vec

Music2U
최도현 곽대훈 김다혜 정도윤 장정주

Index

01 주제선정배경
02 데이터 수집
03 데이터 전처리
04 데이터 분석
05 분석 결과
06 제언

주제
선정 배경
01

기존 음악 추천 프로그램의 한계

좋다고 말해

아티스트 봄빨간사춘기
앨범 Full Album RED PLANET `Hidden Track`
발매일 2016.12.21
장르 Folk

공유하기

스타일이 유사한 인기곡

NO	곡명	아티스트	
1	냉탕에 상어 (Feat. 블랙넛)	슈퍼비 (Superbee)	
2	The Time Goes On	BewhY (비와이)	
3	안아줘	정준일	
4	스토커	10cm	3.0
5	봄이 좋냐??	10cm	3.2

전체선택 | 듣기 | 다운 | 담기 | 선물

기존에 존재하는 음악 추천 프로그램은
주로 유명하거나 우리가 이미 알고있는 노래를
추천해주는데 그치고 있다.

이러한 추천 시스템은
실용성과 만족도 하락의 문제가 존재한다.

진행 과정

데이터 수집 - 수집 데이터 정제 - 데이터 분석 - 추천시스템 구현 - 결론 및 제언

진행 과정

데이터 수집 - 수집 데이터 정제 - 데이터 분석 - 추천시스템 구현 - 결론 및 제언

→ Word2vec : 노래를 벡터 공간에 좌표로 표현

→ LDA: 가사 내용 유사도 측정

→ Daisy : 장르,발매일 반영하여 노래간 유사도 측정

데이터 수집

괜찮은 노래 추천 시스템을 만들고 싶다.

어떻게 하면 내가 좋아하는 노래와 **비슷하면서도 좋은** 노래를 추천받을 수 있을까?

어떤 데이터를 활용해야할까?

데이터 수집

Step1

실제 유저들의 재생목록에 **같이, 자주** 등장하는 곡이 **비슷하고 좋은 곡**일 확률이 높다.

But 우리는 실제 한국 유저들의 재생목록을 구할 수 없다.

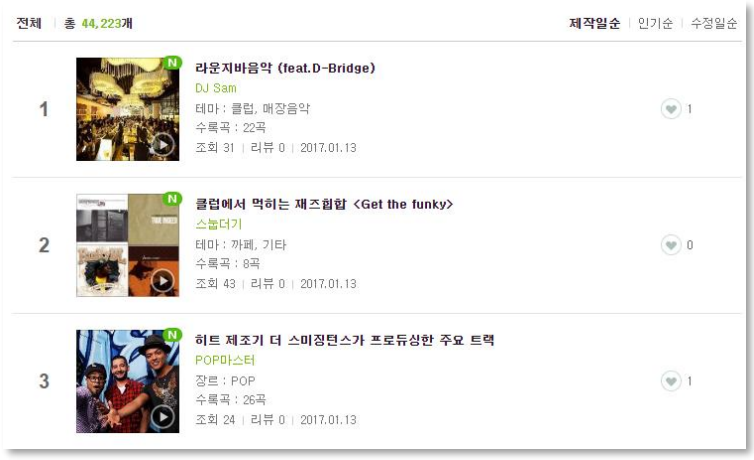
→ **대형 음원사이트의 '유저들의 노래추천목록'을 활용하자!**

Step2

아무도 추천하지 않는 곡보다는 10명이 추천하는 곡이 좋을 것이다.

10명보다는 100명, 1000명이 추천하는 곡이 **더 좋은 추천 곡**일 확률이 높다.

→ **대량의 유저 노래추천목록을 통해 객관성과 신뢰성을 높이자!**



이용 유저가 **2800만명**에 달하는
음원 사이트 Melon에서 정보를 수집하였다.

멜론DJ플레이리스트의
좋아요 표시를 통해 인기도를 확인할 수 있다.

DJ플레이리스트마다
테마가 정해져 있으므로,
어느 정도 노래의 그룹이 분류 되어 있다.

크롤링을 통해 수집한 데이터

① 플레이리스트 정보

playlistid	title	like	songs	vw	date	themalk	트로트	월드뮤직
101571597	짱짱 좋은	26416	131	1364153	2016.12.29	랩/힙합	0	0
101590319	감성이 터	22880	168	840317	2015.07.29	휴식/힐링	0	0
100066911	♪중독성강	20891	75	718353	2013.06.23	까페	0	0
401119058	신비롭거나	19665	299	2186268	2016.07.21	비오는 날	0	0
101606592	내가 까페	18000	90	984656	2014.04.24	휴식/힐링	0	0
101516705	★20대 30	17908	200	1215553	2013.11.10	발라드	0	0
403029596	시험기간	17528	76	814774	2015.09.18	휴식/힐링	0	0
400720502	혼자만 알	16064	320	2488121	2016.08.04	휴식/힐링	0	0
101589497	좋은 노래	15561	200	678480	2016.03.11	발라드	0	0
100135238	언제 들어	15185	70	527579	2011.01.19	R&B/Soul	0	0
401508804	여자들 취	14663	231	2585045	2017.01.02	사랑/설렘	0	0
100098486	가슴을 울	14499	140	481219	2014.05.08	발라드	0	0
407216476	공부 전용	14432	118	677699	2016.12.27	휴식/힐링	0	0
101614141	심판에게	14021	27	682399	2014.02.21	이별/슬픔	0	0
404191009	편하게 듣	13957	180	1351247	2016.08.15	까페	0	0
100115799	카페에서	13669	119	196389	2016.10.05	까페, 기타	0	0
101566985	사랑했었던	13409	397	698269	2015.10.29	록/메탈	0	0
405487309	노래방가면	12633	125	1610227	2016.07.25	스트레스	0	0
402080228	두근두근	11839	218	1928793	2015.10.09	사랑/설렘	0	0
404084040	스튜디오	11608	40	621949	2015.03.19	추억/회상	0	0
400970421	새벽에 들	11435	304	1910646	2016.04.09	밤/새벽	0	0
404337161	아침을 깨	11129	150	1178394	2016.08.15	운동할 때	0	0
407163138	혼자 생각	10549	46	619155	2015.04.22	잔잔한	0	0

...

크롤링을 통해 수집한 데이터

① 플레이리스트 정보

playlistid	title	like	songs	vw	date	themalk	트로트	월드뮤직
101571597	짱짱 좋은	26416	131	1364153	2016.12.29	랩/힙합	0	0
101590319	감성이 터	22880	168	840317	2015.07.29	휴식/힐링	0	0
100066911	♪중독성강	20891	75	718353	2013.06.23	까페	0	0
401119058	신비롭거나	19665	299	2186268	2016.07.21	비오는 날	0	0
101606592	내가 까페	18000	90	984656	2014.04.24	휴식/힐링	0	0
101516705	★20대 30	17908	200	1215553	2013.11.10	발라드	0	0
403029596	시험기간	17528	76	814774	2015.09.18	휴식/힐링	0	0
400720502	혼자만 알	16064	320	2488121	2016.08.04	휴식/힐링	0	0
101589497	좋은 노래	15561	200	678480	2016.03.11	발라드	0	0
100135238	언제 들어	15185	70	527579	2011.01.19	R&B/Soul	0	0
401508804	여자들 취	14663	231	2585045	2017.01.02	사랑/설렘	0	0
100098486	가슴을 울	14499	140	481	2014.05.08	발라드	0	0
407216476	공부 전용	14432	118	677	2016.12.27	휴식/힐링	0	0
101614141	심판에게	14021	27	682	2014.02.21	이별	0	0
404191009	편하게 듣	13957	180	1351	2016.08.15	까페	0	0
100115799	카페에서	13669	119	196	2016.	[illegible]	0	0
101566985	사랑했었던	13409	397	698	2015.10.2	이별/재결	0	0
405487309	노래방가면	12633	125	1610	2016.07.25	스트레	0	0
402080228	두근두근	11839	218	1928	2015.10.09	사랑/	0	0
404084040	스튜디오	11608	40	621	2015.03.19	추억/회상	0	0
400970421	새벽에 들	11435	304	1910	2016.04.09	밤/새벽	0	0
404337161	아침을 깨	11129	150	1178394	2016.08.15	운동할 때	0	0
407163138	혼자 생각	10549	46	619155	2015.04.22	잔잔한	0	0

음악사이트 멜론에서 **DJ플레이리스트**를 활용

아티스트, 곡 제목, 발매일, 장르, 가사, 좋아요 개수 수집

인기순으로 3만개의 DJ플레이리스트 크롤링

크롤링을 통해 수집한 데이터 ② 노래사전

songid	title	artist	date	genre	like	lyric
30007147	Someday	BewhY (비와이)	2016.09.27	Rap/Hip-h	5113	Someday
2785554	어느 비오	비와이패밀리	2010.08.02	Rap/Hip-h	1224	날 떠나지
8271314	쌈박자 (Xa	BewhY (비와이),	2016.07.16	Rap/Hip-h	9989	쌈박자 쌈
7858153	In Trinity	BewhY (비와이)	2015.09.18	Rap/Hip-h	11153	B E W H Y
8261338	Day Day (I	BewhY (비와이)	2016.07.09	Rap/Hip-h	81372	한 번 돌아
5620258	Yelloism	BewhY (비와이)	2015.03.10	Rap/Hip-h	2267	내 피부가
8131716	Shalom	BewhY (비와이)	2016.04.06	Rap/Hip-h	13098	나는 할 필
5620263	몽상	BewhY (비와이)	2015.03.10	Rap/Hip-h	2614	아침엔 알
8283384	Beside Me	코드쿤스트, Bewh	2016.07.26	Rap/Hip-h	8591	Don't neve
5620257	중2병 (Fea	BewhY (비와이)	2015.03.10	Rap/Hip-h	20943	B to the e
7894707	얼어	VASCO (바스코),	2015.10.22	Rap/Hip-h	4153	내가 보이
8252408	Forever (P	BewhY (비와이)	2016.07.02	Rap/Hip-h	77515	래퍼 딱지
5620259	자화상	BewhY (비와이)	2015.03.10	Rap/Hip-h	6374	자화상 대
8271317	자화상 pt.	BewhY (비와이)	2016.07.16	Rap/Hip-h	10224	래퍼 중 제
5620261	No Sky Is	BewhY (비와이)	2015.03.10	Rap/Hip-h	4350	What up Y
8308231	puzzle	씨잼 (C Jamm), Be	2016.08.11	Rap/Hip-h	71986	I'll Take 1
4770079	Swimming	BewhY (비와이)	2014.07.28	Rap/Hip-h	3221	내가 노는

크롤링을 통해 수집한 데이터

② 노래사전

songid	title	artist	date	genre	like	lyric
30007147	Someday	BewhY (비와이)	2016.09.27	Rap/Hip-h	5113	Someday
2785554	어느 비오	비와이패밀리	2010.08.02	Rap/Hip-h	1224	날 떠나지
8271314	쌈박자 (Xa	BewhY (비와이),	2016.07.16	Rap/Hip-h	9989	쌈박자 쌈
7858153	In Trinity	BewhY (비와이)	2015.09.18	Rap/Hip-h	11153	B E W H Y
8261338	Day Day (I	BewhY (비와이)	2016.07.09	Rap/Hip-h	81372	한 번 돌아
5620258	Yelloism	BewhY (비와이)	2015.03.10	Rap/Hip-h	2267	내 피부가
8131716	Shalom	BewhY (비와이)	2016.04.06	Rap/Hip-h	13098	나는 할 필
5620263	몽상	BewhY (비와이)	2015.03.10	Rap/Hip-h	2614	아침엔 알
8283384	Beside Me	코드쿤스트, Bewh	2016.07.26	Rap/Hip-h	8591	Don't nev
5620257	중2병 (Fea	BewhY (비와이)	2015.03.10	Rap/Hip-h	20943	B to the e
7894707	얼어	VASCO (바스코),	2015.10.22	Rap/Hip-h		
8252408	Forever (P	BewhY (비와이)	2016.07.02	Rap/Hip-h	77515	래퍼 딱지
5620259	자화상	BewhY (비와이)	2015.03.10	Rap/Hip-h		
8271317	자화상 pt.	BewhY (비와이)	2016.07.16	Rap/Hip-h	10224	래퍼 좀 재
5620261	No Sky Is	BewhY (비와이)	2015.03.10	Rap/Hip-h	4350	What
8308231	puzzle	씨잼 (C Jamm), Be	2016.08.11	Rap/Hip-h	71986	I'll Take 1
4770079	Swimming	BewhY (비와이)	2014.07.28	Rap/Hip-h	3221	내가 노는

개별 노래의 정보를 멜론에서 크롤링

제목,가수, 발매일, 장르, 좋아요 수, 가사 수집

전체 **38**만 곡

크롤링을 통해 수집한 데이터

② 노래사전

songid	title	artist	date	genre	like	lyric
30007147	Someday	BewhY (비와이)	2016.09.27	Rap/Hip-h	5113	Someday
2785554	어느 비오	비와이패밀리	2010.08.02	Rap/Hip-h	1224	날 떠나지
8271314	쌈박자 (Xa	BewhY (비와이), A	2016.07.16	Rap/Hip-h	9989	쌈박자 쌈
7858153	In Trinity	BewhY (비와이)	2015.09.18	Rap/Hip-h	11153	B E W H Y
8261338	Day Day (I	BewhY (비와이)	2016.07.09	Rap/Hip-h	81372	한 번 돌아
5620258	Yelloism	BewhY (비와이)	2015.03.10	Rap/Hip-h	2267	내 피부가
8131716	Shalom	BewhY (비와이)	2016.04.06	Rap/Hip-h	13098	나는 할 필
5620263	몽상	BewhY (비와이)	2015.03.10	Rap/Hip-h	2614	아침엔 알
8283384	Beside Me	코드쿤스트, Bewh	2016.07.26	Rap/Hip-h	8591	Don't nev
5620257	중2병 (Fea	BewhY (비와이)	2015.03.10	Rap/Hip-h	20943	B to the e
7894707	얼어	VASCO (바스코), A	2015.10.22	Rap/Hip-h		
8252408	Forever (P	BewhY (비와이)	2016.07.02	Rap/Hip-h	77515	래퍼 딱지
5620259	자화상	BewhY (비와이)	2015.03.10	Rap/Hip-h		
8271317	자화상 pt.	BewhY (비와이)	2016.07.16	Rap/Hip-h	10224	래퍼 중
5620261	No Sky Is	BewhY (비와이)	2015.03.10	Rap/Hip-h	4350	What's
8308231	puzzle	씨잼 (C Jamm), Be	2016.08.11	Rap/Hip-h	71986	I'll Take 1
4770079	Swimming	BewhY (비와이)	2014.07.28	Rap/Hip-h	3221	내가 노는

전체 **38**만 곡 중 플레이리스트에서
5번 이상 나온 노래만 채택하여 총 **8**만 곡을
대상으로 분석

데이터 전처리

03

1

노래의 장르 정리

2

결측치 제거

3

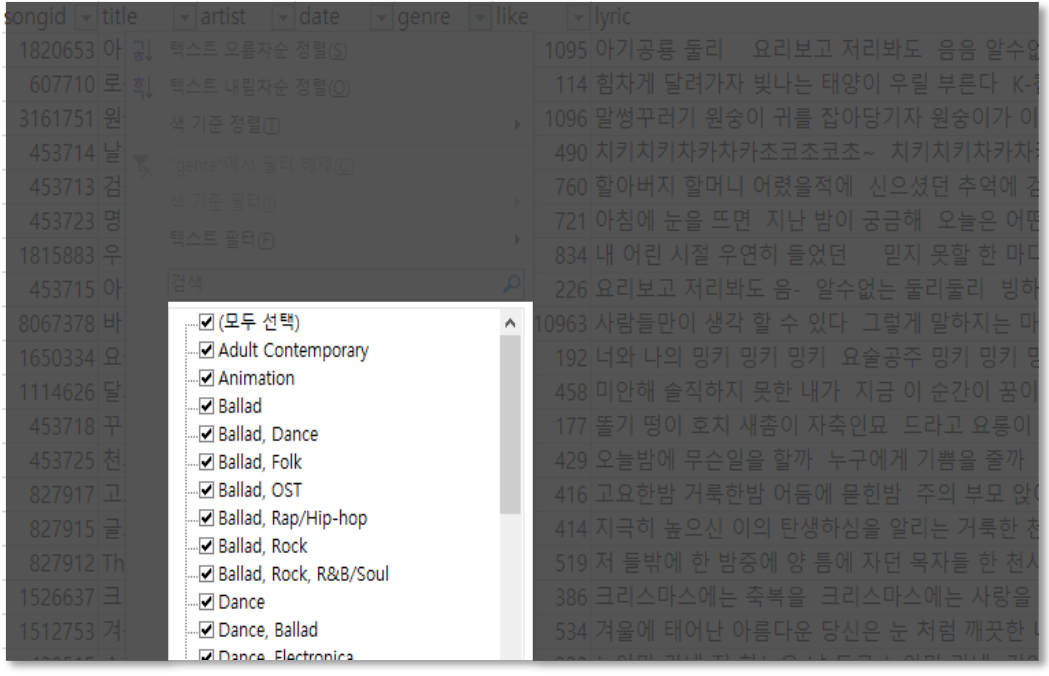
외국 노래 제거

4

새로운 변수 추가

1 노래의 장르 정리

약 50개 종류의 장르를 통합, 삭제하여 11개로 정리하였다.



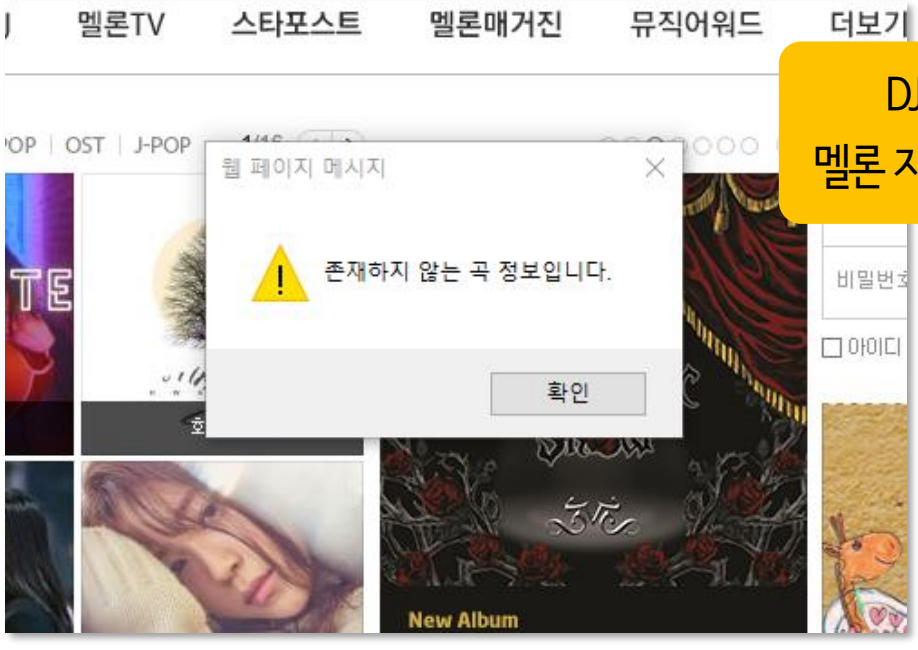
플레이리스트의 주요 장르를 제외한 나머지 장르들은 **통합** or **삭제**하였다.

통합 - (Korean movie+Drama=OST)
삭제 - (Pop, 클래식, 찬송가 등)

여러 장르가 중복된 경우 빈도수를 셀 때 포함될 수 있도록 콤마(,)로 구분하여 넣어 주었다.

2 결측치 처리

① 삭제된 노래



DJ 플레이리스트에 포함된 곡 중
멜론 자체에서 삭제된 노래를 제거하였다.

2 결측치 처리
②기타

songid	title	artist	date	genre	like	lyric
1002914	Harder, Be	Daft Punk	2001.03.13	Electronica	10182	Work It Make It Do It Makes Us Harder Better Faster Strong
1002916	Nightvisio	Daft Punk	2001.03.13	Electronica	250	[가사 준비중] 멜론 회원 여러분! 가사 등록을 기다리고 있어요.
1002919	Something	Daft Punk	2001.03.13	Electronica	17074	It might not be the right time I might not be the right one But
1002920	Voyager	Daft Punk	2001.03.13	Electronica	441	[가사 준비중] 멜론 회원 여러분! 가사 등록을 기다리고 있어요.
1002921	Veridis Qu	Daft Punk	2001.03.13	Electronica	323	[가사 준비중] 멜론 회원 여러분! 가사 등록을 기다리고 있어요.
1002923	Face To Fa	Daft Punk	2001.03.13	Electronica	1079	What's going on Could this be my understanding It's not your
1002986	Beethover	곽윤찬	2005.10.20	Jazz	146	[가사 준비중] 멜론 회원 여러분! 가사 등록을 기다리고 있어요.
1003011	Call Me	웅산	2005.11.07	Jazz	89	웅산 The Blues - Call Me Call me.. When you feel so blue
1003049	못다핀 꽃	김수철	1983.08.15	Rock	1698	언제 가셨는데 안오시나 한잎두고 가신님아 가지위에 눈물 적셔늘
1003056	내일	김수철	1983.08.15	Rock	1047	스쳐가는 은빛 사연들이 밤하늘에 가득차고 풀나무에 맺힌 이슬처
1003063	Maybe I	Cooly`s H	2004.11.23	Electronica	764	I didn't think I would get such an inspiration
1003072	젊은 그대	김수철	1984.10.01	Rock	224	거칠은 벌판으로 달려가자 젊음의 태양을 마시자 보석보다 찬란한
1003075	나도야 간	김수철	1984.10.01	Rock	372	봄이 오는 캠퍼스 잔디밭에 팔베개를 하고 누워 편지를 쓰네 노랑
1003080	앵무새	하울	2005.11.15	Ballad	7178	또 어제처럼 다시 그립습니다 보고 싶은 맘 줄지도 않는지 자꾸만
1003293	일급 비밀	소방차	1988.01.01	Dance	263	눈을 감아도 소용없어 귀를 막아도 소용없어 지금 그녀가 내 앞에

2 결측치 처리
②기타

songid	title	artist	date	genre	like	lyric
1002914	Harder, Be	Daft Punk	2001.03.13	Electronica	10182	Work It Make It Do It Makes Us Harder Better Faster Strong
1002916	Nightvisio	Daft Punk	2001.03.13	Electronica	250	[가사 준비중] 멜론 회원 여러분! 가사 등록을 기다리고 있어요.
1002919	Something	Daft Punk	2001.03.13	Electronica	17074	It might not be the right time I might not be the right one But
1002920	Voyager	Daft Punk	2001.03.13	Electronica	441	[가사 준비중] 멜론 회원 여러분! 가사 등록을 기다리고 있어요.
1002921	Veridis Qu	Daft Punk	2001.03.13	Electronica	323	[가사 준비중] 멜론 회원 여러분! 가사 등록을 기다리고 있어요.
1002923	Face To Fa	Daft Punk	2001.03.13	Electronica	1079	What's going on Could this be my understanding It's not your
1002986	Beethover	곽윤찬	2005.10.20	Jazz	146	[가사 준비중] 멜론 회원 여러분! 가사 등록을 기다리고 있어요.
1003011	Call Me	웅산	2005.11.07	Jazz	89	웅산 The Blues - Call Me Call me.. When you feel so blue
1003049	못다핀 꽃	김수철	1983.08.15	Rock	1698	언제 가셨는데 안오시나 한잎두고 가신님아 가지위에 눈물 적셔
1003056	내일	김수철	1983.08.15	Rock	1047	스쳐가는 은빛 사연들이 밤하늘에 가득차고 풀나무에 맺힌 이슬
1003063	Maybe I	Cooly`s H	2004.11.23	Electronica	764	I didn't think I would get such an inspiration
1003072	젊은 그대	김수철	1984.10.01	Rock	224	거칠은 벌판으로 달려가자 젊음의 태양을 마시자 보석보다 찬란
1003075	나도야 간	김수철	1984.10.01	Rock	372	봄이 오는 캠퍼스 잔디밭에 팔베개를 하고 누워 편지를 쓰네 노
1003080	앵무새	하울	2005.11.15	Ballad	7178	또 어제처럼 다시 그립습니다 보고 싶은 맘 줄지도 않는지 자꾸
1003293	일급 비밀	소방				

NA가 존재하는 경우, 가사 없는 노래, 가사가 너무 짧은 노래를 제거해주었다.

3 외국 노래 제거

한국 노래 맞춤 추천을 위해 해외 노래를 제거하였다.

3 외국 노래 제거

한국 노래 맞춤 추천을 위해 해외 노래를 제거하였다.

➔ 클래식, POP, 일렉트로니카, J-POP, 월드뮤직, 뉴에이지 제거해 주었으며, 이외의 걸러지지 않은 영어, 일본어, 중국어 노래는 수작업으로 제거

4 새로운 변수 추가

① 테마

랩.힙합 / 휴식.힐링 / 까페 / … / 감성 / 트로트로 이루어진 총 29가지의 테마

해당 곡이 4만여개의 DJ플레이리스트들 중 각각의 테마에 몇 번 포함되었는지 비율로 나타내었다.

<어반자카파 - 널 사랑하지 않아>

*DJ 플레이리스트 총 407번 등장

휴식.힐링 29번 $\frac{29}{407} = 0.071253$

감성 3번 $\frac{3}{407} = 0.007371$

songid	title	artist	랩.힙합	휴식.힐링	…	감성
8199190	널 사랑하지 않아	어반자카파	0	0.071253	…	0.007371

4 새로운 변수 추가

② season(계절) / ③ age(발매연도)

제공된 앨범 발매일을 통해 계절과 발매연도 변수를 생성한다.

season(계절)

각 앨범의 발매월을 기준으로 계절 구분

봄	3월, 4월, 5월
여름	6월, 7월, 8월
가을	9월, 10월, 11월
겨울	12월 ,1월, 2월
unknown	기타

age(발매연도)

각 앨범의 발매연도를10년 단위로 범주화

1990년도 이전
1990년대
2000년대
2010년 이후

4 새로운 변수 추가

② season(계절) / ③ age(발매연도)

제공된 앨범 발매일을 통해 계절과 발매연도 변수를 생성한다.

Ex)

<어반자카파 - 널 사랑하지 않아>

발매일: 2016.05.27

➔ Season = 봄

➔ Age = 2010년 이후

최종 데이터

전체 8만곡 중 결측치가 존재하거나 외국 노래를 제외하였을 때 대략 **36000곡**이 남았다.

변수: songid / title / artist / genre / like / lyric / 29가지테마 / season / age

songid	title	artist	genre	like	lyric	랩.힙합	트로트	season	age
2328156	왜 나만 아프죠	아이비	Ballad	953	나를 미워	0	0	가을	2000년대
1620554	날 위한 이별	김혜림	Ballad	3346	난 알고 있	0	0	가을	1990년대
4800348	잘할게요	브라운아이드	Ballad	765	난 지금 이	0	0	여름	2010년이후
1314777	그래도 사랑하니까	박정은	Ballad	122	내게 상처	0	0	여름	2000년대
4027922	사랑하고 싶어서	허각	Ballad	10812	노을 빛 구	0.04	0	겨울	2010년이후
487287	Please	이기찬	Ballad	170	누군가 지	0	0	가을	2000년대
47133	가로수 그늘아래 서	이문세	Ballad	2094	라일락 꽃	0.061	0	가을	1990년 이전
91403	방황	이승철	Ballad	483	매일 신문	0	0	가을	1990년대
932539	당신은 내게 그런	소냐	Ballad	72	머릴 감지	0	0	가을	2000년대
8306285	미워하는 마음 만	지아	Ballad	741	미워하는	0	0	여름	2010년이후
1923447	사랑은.. 항상...	더 원	Ballad	205	사랑은 항	0	0	가을	2000년대
1923575	두 가지 일	먼데이 키즈	Ballad	994	생각나는	0	0	가을	2000년대
537609	Silent Eyes	이수영	Ballad	96	안개속에 서	0	0	가을	2000년대
189692	슬픈 그림같은 사	이상우	Ballad	200	안녕이라는	0	0	여름	1990년 이전
462462	슬픈 그림같은 사	이상우	Ballad	258	안녕이라는	0	0	봄	2000년대
52159	슬픈 그림같은 사	이상우	Ballad	956	안녕이라는	0	0.027778	봄	1990년 이전
4626022	누른다	김꽃	Ballad	591	억지로 누	0	0	봄	2010년이후

데이터
분석
04

1
Word2vec

2
LDA

3
Daisy

1 Word2Vec
개념, 프로젝트에서의 적용

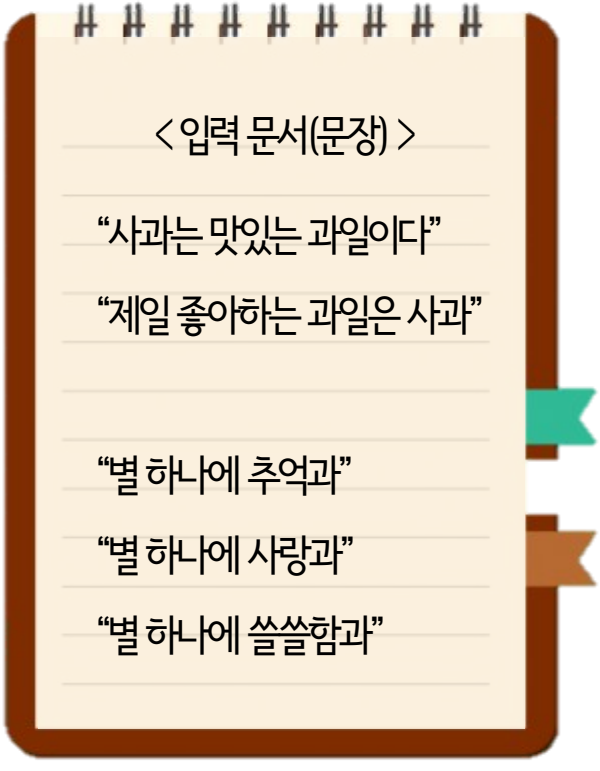
Word2vec $\arg\max_{\theta} \prod_{(w,c)\in D} p(c|w;\theta)$

Word2vec이란 입력한 텍스트 문서에서 각각의 단어(w)가 문맥(c)상에서 결합될 확률이 최대가 되도록, **단어 w의 벡터 값을 학습**하는 알고리즘.

비슷한 의미의 단어들은 **문서상 가까운 곳**에서 출현할 확률이 높다.
그래서 많은 양의 텍스트 문서를 학습해가면서
비슷한 단어들은 점차 가까운 벡터(좌표)를 가지게 된다.

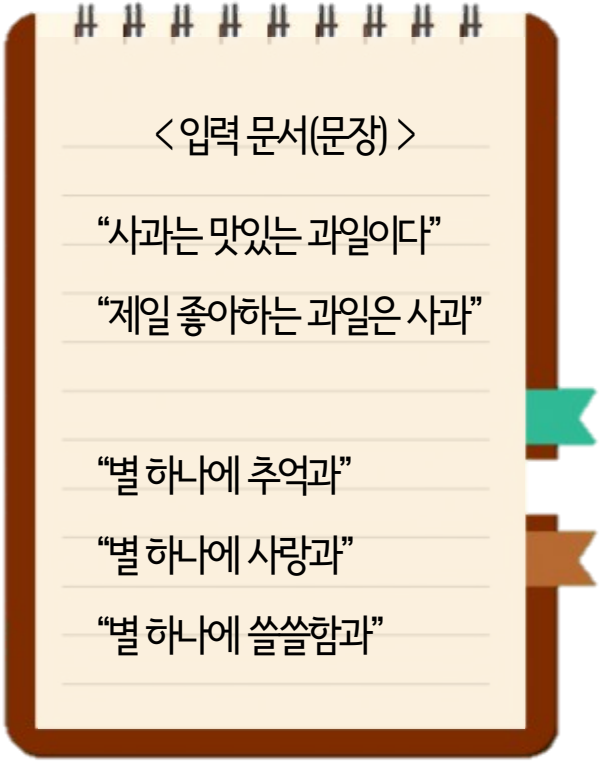
1 Word2Vec

개념, 프로젝트에서의 적용

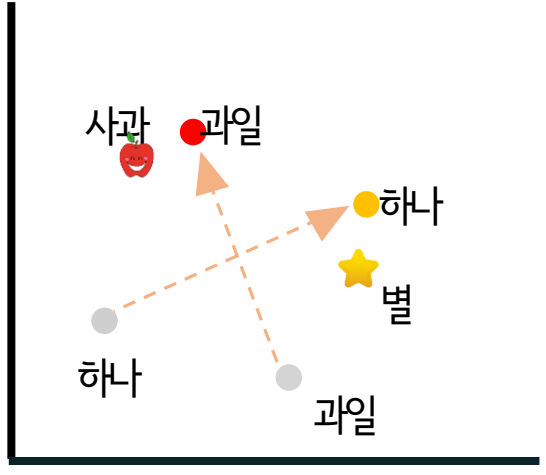


1 Word2Vec

개념, 프로젝트에서의 적용

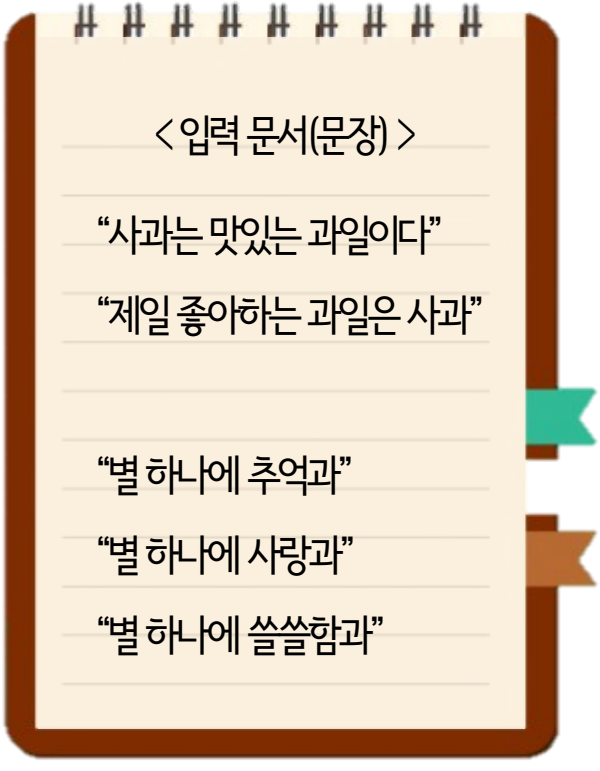


<학습 후>

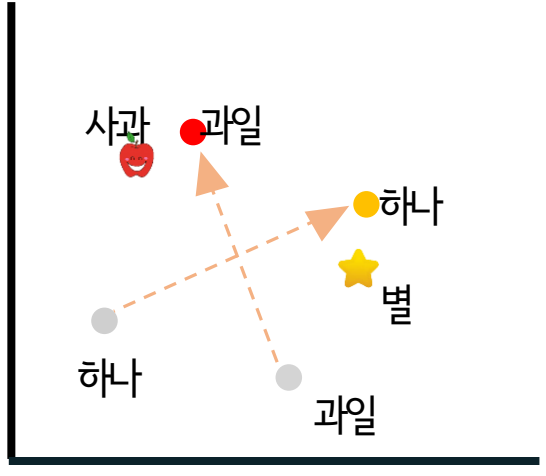


1 Word2Vec

개념, 프로젝트에서의 적용



<학습 후>



➔ Input은 문장이고 Output은 단어들의 벡터 값 좌표

1 Word2Vec

개념, 프로젝트에서의 적용

단어

단어1, 단어2, ...

EX)가방,가수

문장

{단어7,단어5,단어3}

EX)나는,학교에,간다

문서

{문장1,문장2,…}

{{나는,학교에,간다},

{오늘은,춥다}…}

1 Word2Vec

개념, 프로젝트에서의 적용

단어	문장	문서
단어1, 단어2, ...	{단어7,단어5,단어3}	{문장1,문장2,…}
EX)가방,가수	EX)나는,학교에,간다	{{나는,학교에,간다},{오늘은,춥다}…}
↓	↓	↓
음악	플레이 리스트	문서
음악1, 음악2, …	{음악3,음악5,음악7..}	{플레이리스트1 플레이리스트2,…}

1 Word2Vec
개념, 프로젝트에서의 적용

< Word2vec을 이용한 분석과정 >

Input은 문장 -> DJ플레이리스트(노래 목록)

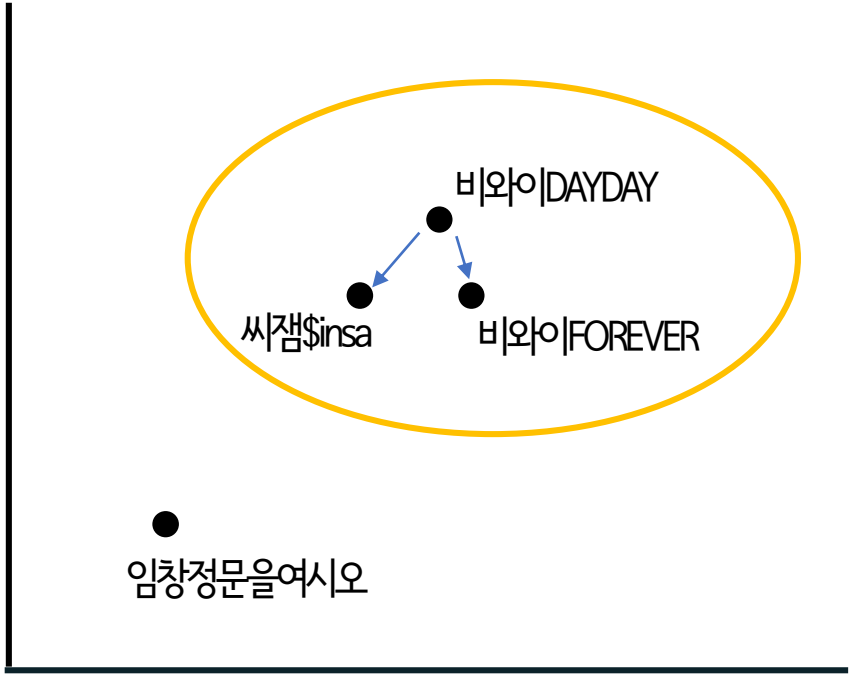
ex) “사과는 과일이다” -> “비와이DAYDAY 비와이FOREVER”

Output은 단어의 벡터 -> 노래의 벡터

ex) 사과 = (1,3) -> 비와이DAYDAY = (2, 2)

1 Word2Vec

개념, 프로젝트에서의 적용



내가 좋아하는 곡을 고르면,
그 곡과 벡터 공간상에서
가장 가까운 노래들을 추천

1 Word2Vec
개념, 프로젝트에서의 적용

한글 문장
word2vec

→ 단어 순서 중요
→ 나는 학교에 간다.
학교에 나는 간다. 의 학습 결과는 다름

플레이리스트(문장)에서 음악ID(단어) 배치는 어떻게 할까?

1 Word2Vec

개념, 프로젝트에서의 적용

EX) 플레이리스트1->{음악1, 음악10, 음악5, 음악3…}

플레이리스트 1-1 ->{음악10,음악1,음악3,음악5…}

플레이리스트 1-2 ->{음악3, 음악1, 음악5, 음악10…}

…

플레이리스트 1-10 ->{음악1,음악5,음악3,음악10…}

단어의 비율은 그대로 유지되며

단어 순서의 중요도는 축소되었다.

3만개 플레이리스트 → 30만개 플레이리스트

2 LDA(Latent Dirichlet Allocation)

노래 가사 키워드 반영

LDA 는

Unsupervised Generative Topic Model의 일종으로

문서를 모델링하는 기법이다.

2 LDA(Latent Dirichlet Allocation)

노래 가사 키워드 반영

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

어떤 문서들을 학습 시키면 문서 내 단어들이
주제나 사용 용도에 따라 **군집화** 됨

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

2 LDA(Latent Dirichlet Allocation)

노래 가사 키워드 반영

LDA 군집화 결과

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]
[1,]	"la"	"널"	"그"	"왜"	"널"	"you"	"i'm"	"난"	"you"	"그대"
[2,]	"up"	"너를"	"이"	"너"	"너를"	"i"	"fuck"	"그"	"love"	"그댈"
[3,]	"it"	"사랑해"	"저"	"좀"	"수"	"the"	"내"	"내"	"oh"	"그대를"
[4,]	"hey"	"너의"	"밤"	"넌"	"다시"	"to"	"난"	"이"	"i"	"그대가"
[5,]	"get"	"내"	"비가"	"안"	"그"	"me"	"a"	"수"	"baby"	"그댄"
[6,]	"na"	"나의"	"하얀"	"니가"	"너"	"and"	"the"	"건"	"my"	"그대의"
[7,]	"go"	"너"	"속에"	"그냥"	"니가"	"my"	"rap"	"나의"	"me"	"말아요"
[8,]	"let's"	"너와"	"나의"	"난"	"왜"	"it"	"shit"	"꿈을"	"girl"	"내"
[9,]	"down"	"너에게"	"함께"	"내가"	"날"	"a"	"돈"	"더"	"u"	"그대는"
[10,]	"on"	"있어"	"너의"	"니"	"난"	"in"	"money"	"않아"	"so"	"사랑"
[11,]	"everybody"	"수"	"바람이"	"좋아"	"눈물이"	"be"	"du"	"내가"	"wanna"	"수"
[12,]	"shake"	"함께"	"다시"	"해"	"내"	"your"	"man"	"있어"	"i'm"	"사랑해요"
[13,]	"dance"	"곁에"	"내"	"너무"	"너의"	"we"	"니"			
[14,]	"party"	"싶어"	"작은"	"나"	"나"	"is"	"u"			
[15,]	"hot"	"난"	"바람"	"몰라"	"또"	"i'm"	"ye			
[16,]	"come"	"영원히"	"우리"	"그래"	"없는"	"on"	"다"			
[17,]	"boom"	"그"	"그대와"	"잘"	"이렇게"	"all"	"like"	"다시"	"know"	"있어요"
[18,]	"rock"	"언제나"	"하늘"	"다"	"제발"	"that"	"안"	"나는"	"be"	"그대만"
[19,]	"da"	"너만"	"눈이"	"돼"	"이젠"	"for"	"we"	"없어"	"bye"	"사랑이"
[20,]	"the"	"내가"	"내리는"	"정말"	"없어"	"of"	"no"	"한"	"no"	"있나요"

36000개의 가사들을 모두 학습 시켜 가사 내 단어들을 군집화 함

<노래 A의 가사>

나는 학교에 간다.
날씨가 춥다.

군집1	군집2	군집3
나는	축구	간다.
나의	학교	춥다.

<노래 A의 가사>

나는 학교에 간다.
날씨가 춥다.

군집1	군집2	군집3
나는	축구	간다.
나의	학교	춥다.

빈도 계산

군집1	군집2	군집3
1	0	1
0	0	1

합(행)

군집1	군집2	군집3
1	0	2

<노래 A의 가사>

나는 학교에 간다.
날씨가 춥다.

군집1	군집2	군집3
나는	축구	간다.
나의	학교	춥다.

빈도 계산

군집1	군집2	군집3
1	1	1
0	0	1

합(행)

군집1	군집2	군집3
1	1	2

비율 계산

군집1	군집2	군집3
1/4	1/4	2/4

3 Daisy
장르,발매일 반영하여 노래간 유사도 측정

songid	title	distance
3599484	바질 (Feat. BrotherSu)	0.10342
4261268	우리는 (Feat. Crybaby, DJ Dopsh)	0.134793
4607133	우리는 (Feat. Crybaby, DJ Dopsh)	0.116452
3829178	내몸이 불 타오르고 있어 (With 최단비)	0.117091
3434861	Hello My Dear	0.157274
3139148	Suga Luv (Valentine Mix) (Feat. 아이유)	0.123016
522852	너에게 쓰는 편지 (Feat. 린)	0.112462
8151868	그냥 다 들어줄게 (Feat. 오상아)	0.086683
3836555	일단은 달달한 노래 (Feat. Gamjay, 임미소)	0.115231
30094293	빨래방 (Feat. BK, 입술세개)	0.132256
3539378	우아한년 2012 (Feat. San-E & Okasian)	0.126633
1570013	내 모든걸 줄게	0.141373

<널 사랑하지 않아와 나머지 곡 간의 거리>

LDA를 통해 만들어진 가사 클러스터 10개 변수와
장르, 계절, 발매연도, 테마를
Daisy함수를 통해 비유사성을 측정한다.

입력된 곡과 나머지 곡들의 거리를 계산해
추천 목록을 한번 더 필터링

변수를 살펴보면,
장르 / 계절 / 발매연도는 **범주형** 변수,
LDA 변수와 테마는 **수치형** 변수에 해당하므로
혼합형 변수의 거리 계산을 위해
Gower방법을 적용하였다.

함께 음악을 들어 볼까요?

Word2Vec만으로 구현한 추천음악 목록

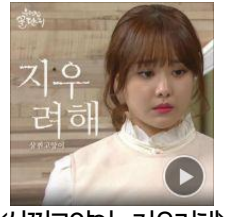
<어반자카파 - 널 사랑하지 않아>

<버나드박 - 하고싶은말>


<김나영 - 어땠을까>


<선인장 - 별이되기전어느날>

<엠씨더맥스-어디에도>
<엠씨더맥스-아스라이>



<살찐고양이 - 지우려해>


<윤시윤 - 사귀고싶어>

<멜로망스 - 봄이되어준그대>

<소방차-어젯밤 이야기>

<동자-브이>

가사,장르,발매일 등을 변수로 생성하여
비유사도를 측정하여 추가로 필터링을 거치면

<버나드박 - 하고싶은말>


<김나영 - 어땠을까>

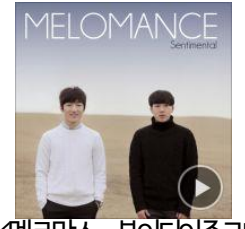

<선인장 - 별이되기전어느날>

<어반자카파 - 널 사랑하지 않아>

<엠씨더맥스-어디에도>
<엠씨더맥스-아스라이>

<살찐고양이 - 지우려해>


<윤시윤 - 사귀고싶어>



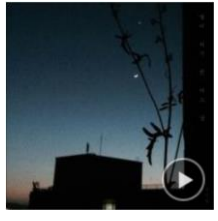
<멜로망스 - 봄이되어준그대>

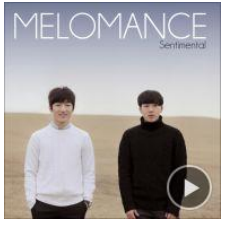
<소방차-어젯밤 이야기>

<동자-브이>

<어반자카파 - 널 사랑하지 않아>

최종 추천리스트



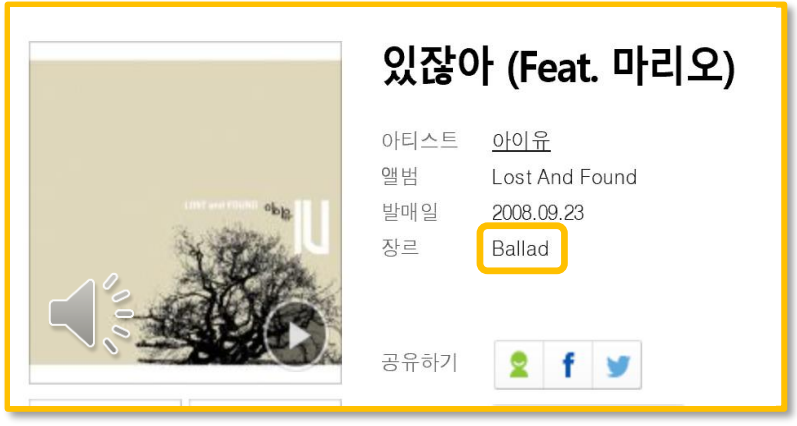


제언

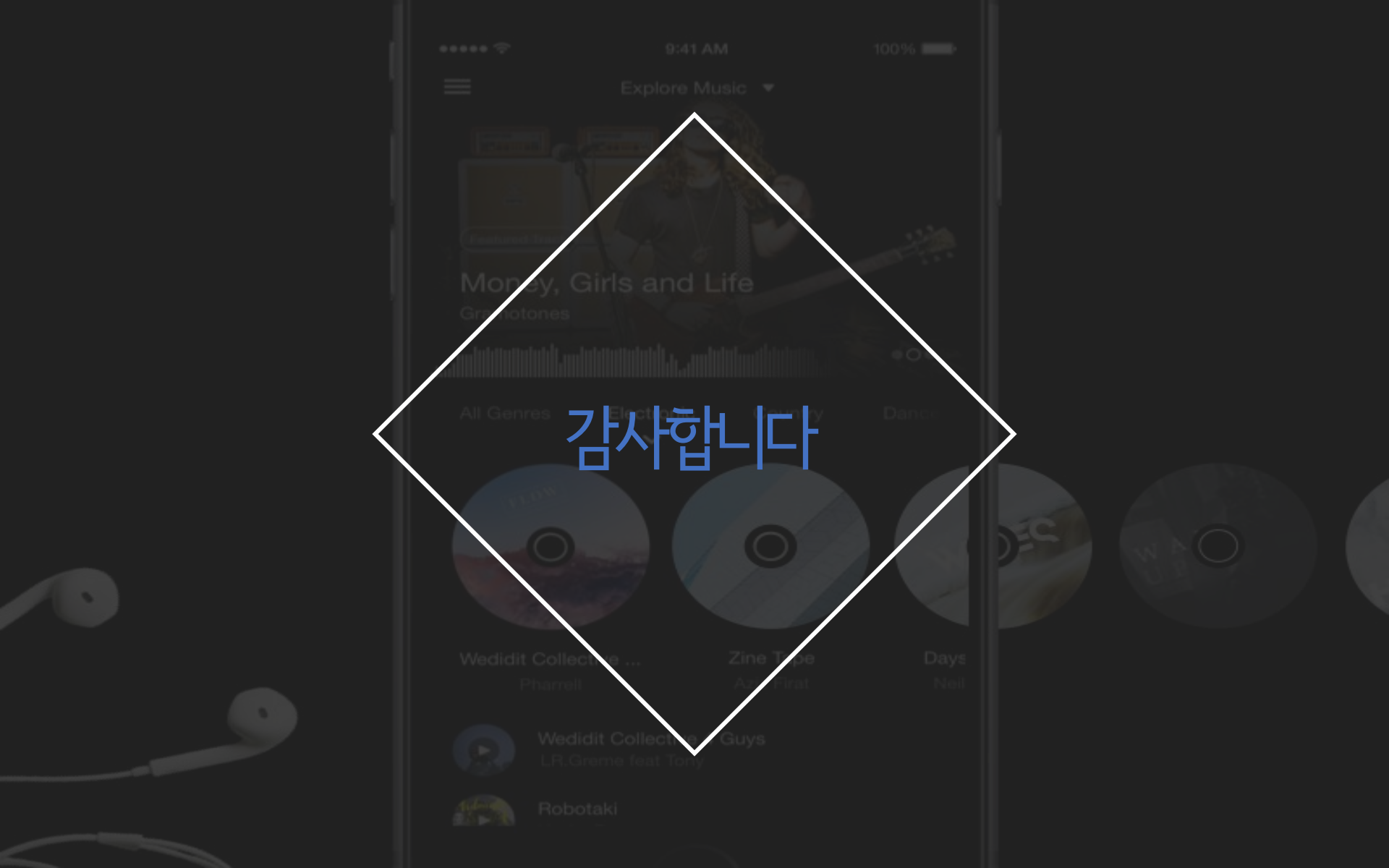
06

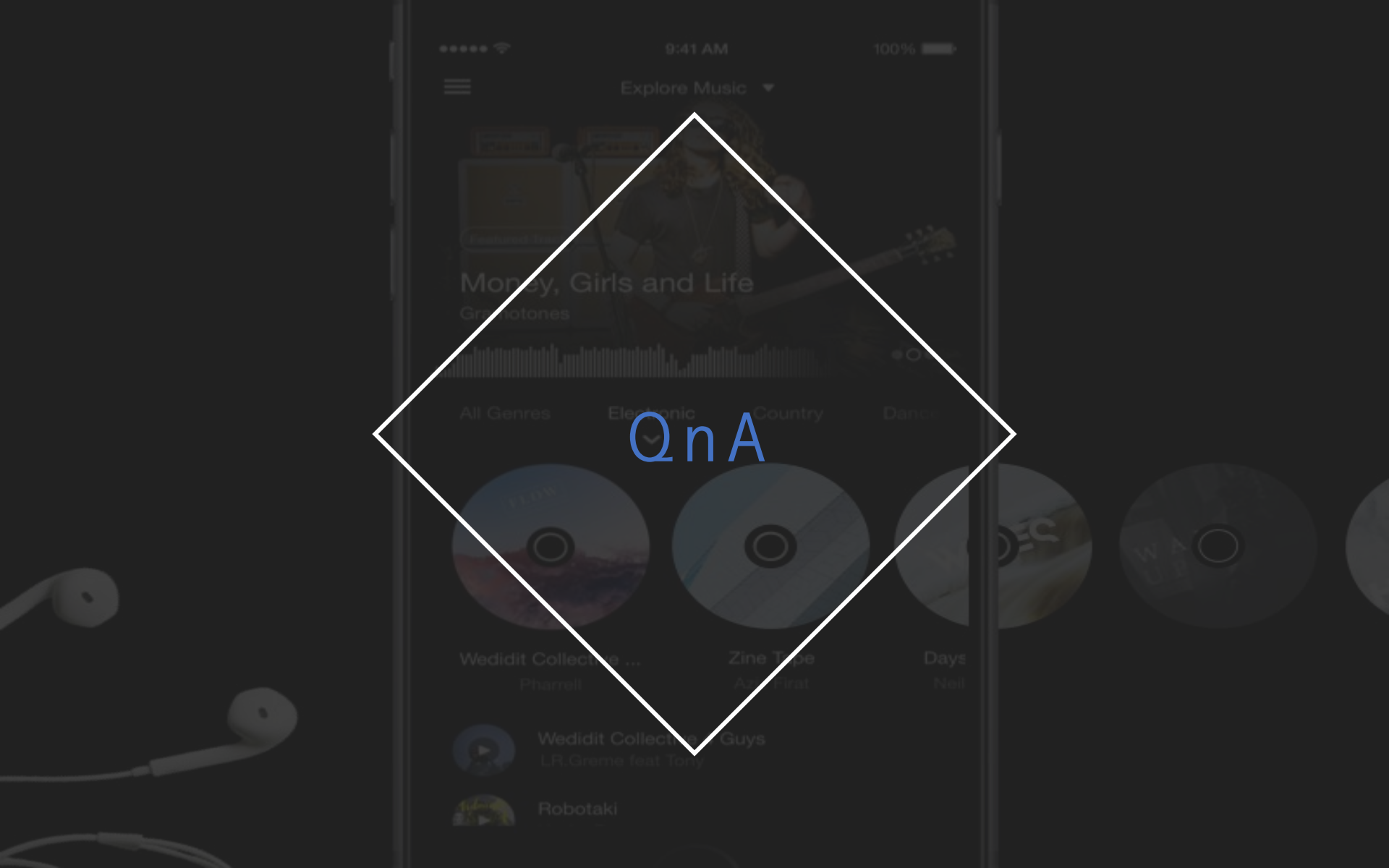
1. 유명하지만 가사가 등록되지 않은 노래, 장르의 구분이 애매하거나 아예 잘못 등록되어 있는 노래 등을 포함하지 못했다.
-> 조금 더 완성도 있는 추천 시스템을 구현함에 있어서 제한

ex) '아이유 - 있잖아' 가 Ballad 장르로 등록



2. 신곡에 대한 추천은 만족도가 떨어질 가능성이 높다.
-> 유저들의 추천데이터를 기반으로 하기 때문에 충분히 쌓여야 퀄리티 높은 추천이 가능


감사합니다


QnA