A young man with dark hair is adjusting a pair of headphones on a young woman with dark hair and bangs. They are in a crowded, dimly lit setting, possibly a concert or a party. The man is looking down at the woman's head, and she is looking forward with a neutral expression. The background is blurred, showing other people in the crowd.

[들려줄게, 너의 노래]

인하대학교 통계학과
빅데이터 자료분석

박정우 장정주

INDEX

01 주제 소개

02 데이터 수집

03 데이터 정제

04 데이터 분석

05 분석 결과

06 한계 및 제언



01 주제 소개





01

주제 소개

kimh**** 댓글모음 >

이국종 박사님! 진정 당신은 **슈바이처**입니다.

2017-11-25 09:13 | 접기요청

답글

👍 13 🗨 0

ninj**** 댓글모음 >

엄마는 뭐하고 있었을까? 10초 동안, .개도 **덜 꼬** 다니면 10초 동안 방치하기 어려운데...

2017-09-30 06:50 | 접기요청

답글 10

👍 527 🗨 18

sb90****

엄마 라는 그년이 **찐빠**대만 고생하셨어요.

2017-09-30 07:34 | 접기요청

답글

👍 34 🗨 1

sksx****

나쁜사람은없어요. 상황에 따라 달라져서그렇지.

2017-11-25 09:10 | 접기요청

답글

👍 6 🗨 0

뉴스 댓글의 문제점

- 오타
- 맞춤법 오기
- 욕설
- 띄어쓰기 불투명



01

주제 소개

‘정규화 된 텍스트’를 이용하기 위해
분석주제 재선정





01

주제 소개

‘나의 플레이리스트’ 기반 음악추천

[특징]

1. 가사의 유사도 반영
2. 혼합변수의 유사성 계산
3. 감상횟수로 가중치 부여

*정규화 된 텍스트 = 노래 가사





01

주제 소개

많이 들은곡

홈 > 마이뮤직 > 감상기록 > 많이 들은곡

▶ 듣기 + 추가 📁 담겨 ⬇ 다운로드

번호	곡정보	감상횟수	더보기
1	매일 듣는 노래 (A Daily Song) 황치열 ▾ Be ordinary	총 81회	⋮
2	빌려줄게 신용재 (포맨) EMPATHY	총 71회	⋮
3	틀리나요 The One (더원) 틀리나요	총 63회	⋮
4	숨쉬는 모든 날 팍키 ▾ 수상한 파트너 OST Part.6 (SBS ...	총 62회	⋮
5	봄날의 소나기 (Paper Um brella) 예성 (YESUNG) Spring Falling - The 2nd Mini Al...	총 56회	⋮
6	나의 사춘기에게 봄방간사춘기 ▾ Red Diary Page.1	총 49회	⋮
7	괜 알꺼야 봄방간사춘기 ▾ Red Diary Page.1	총 48회	⋮
8	고쳐주세요 봄방간사춘기 ▾ Red Diary Page.1	총 47회	⋮
9	한 편의 너 강다니엘 & 도겸 (SEVENTEEN) 한 편의 너	총 47회	⋮
10	무슨 말이 더 필요해	총 47회	⋮

[Genie - 마이 뮤직]

내가 들은 노래에 대한 정보
+
각 노래의 감상횟수



02 데이터 수집





02

데이터 수집

Using Python Selenium

01 지니 [나의 플레이리스트] *genie*

	ID	Title	Artist	Album	Cnt
	1	1 매일 듣는	황치열	Be ordinary	81
	2	2 빌려줄게	신용재 (포	EMPATHY	71
	3	3 들리나요	The One (들리나요	63
	4	4 숨쉬는 모	범키	수상한 파	62
	5	5 봄날의 소	예성 (YES	Spring Fall	56
	6	6 나의 사춘	볼빨간사춘	Red Diary	49
			.		
			.		
			.		
			.		
			.		
	285	95 무제(無題)	G-DRAGO	권지용	17
	286	96 I Love You	수지	당신이 잠	17
	287	97 All I Wann	박재범	EVERYTHI	17
	288	98 봄날	방탄소년단	YOU NEVER	17
	289	99 어디에도	엠씨더맥스	pathos	17
	290	100 어떤 오후	소녀시대	Lion Heart	17

총 290 곡
수집





02

데이터 수집

Using Python Selenium

02 멜론 [최신곡 - 국내] Melon

title	artist	album	date	genre	flac	like	reply	lyric	lyricist	composer	arranger
플라시보 (슬리피	플라시보	플라시보	2017.11.18	Rap / Hip	Flac 16bit	469명	15개	언젠간 디	슬리피, Ha	ASSBRASS	ASSBRASS
삼킨다 (Sc반하나	우리	우리	2017.11.18	Ballad	Flac 16bit	642명	5개	안녕 오랜	Mr. Black	피아노맨 (피아노맨 (김세정)	
Perfect Lif	에브리 싱	고백부부 (2017.11.18	Drama	Flac 16/24	315명	3개	오 you are	문성남 (에	문성남 (에	문성남 (에브리싱글더
끝사랑	다비치	불후의 명	2017.11.18	Ballad		327명	3개	나는 다시	윤사라	윤일상	돈 스파이크
한잔할까	문명진	한잔할까	2017.11.18	R&B / Soul		651명	5개	한잔할까	문명진, Grc	문명진, Grc	Groovin
집으로 가	곽은기 (은	변함없이	2017.11.18	Folk	Flac 16/24	76명	1개	어제 쌓인	곽은기 (은	곽은기 (은	곽은기 (은훤)
결국 너야	더필름	결국 너야	2017.11.18	Ballad	Flac 16bit	248명	2개	피곤한 하	더필름	더필름	더필름, 임정규
오늘은	넬 (NELL), NELL X GF		2017.11.17	Rock	Flac 16bit	8,224명	83개	잘 모르겠	김종완 (NI	김종완 (NI	넬 (NELL), 그루비룸 (C
별이될게	DK, 지아	별이될게	2017.11.17	Ballad	Flac 16/24	2,641명	16개	널 사랑한	검은띠	검은띠	성규호
피카부 (Pe	Red Velve	Perfect Ve	2017.11.17	Dance	Flac 16bit	43,490명	622개	Uhm yeah	kenzie	문샤인, Caz	문샤인
Ms.808 (Fe	Viann, Kh	Ms.808	2017.11.18	Rap / Hip	Flac 16bit	101명	0개	처음처럼	Khundi Pa	Viann, SUN	Viann
사랑 그 쓸	최백호	더 마스터	2017.11.18	Drama	Flac 16bit	35명	0개	다시 또 누	양희은	이병우	알고보니 혼수상태, 김
Brown lip	Knave (네	Brown lip	2017.11.18	R&B / So	Flac 16bit	61명	2개	오늘 일은	Knave (네	M.Fasol, Kr	M.Fasol, Knave (네이브
날 두고 떠	송홍섭	송홍섭 양	2017.11.18	Electronic	Flac 16bit	7명	0개	그대 눈을	송홍섭	송홍섭	송홍섭
Crazy	Dok2	CRAZY	2017.11.17	Rap / Hip	Flac 16bit	1,312명	19개	난 돈 원하	Dok2	GRAY (그	GRAY (그레이)
제발 우리	태사비에	보그맘 OS	2017.11.17	Drama	Flac 16/24	188명	2개	시간은 밤	김성채, cin	손이삭	손이삭
이사 전 날	한소아	널 헤는 밤	2017.11.17	Ballad	Flac 16/24	172명	0개	아무 것도	JQ, 이지혜	강균성 (노	박철호, 임재신

총 45400 곡
수집





02

데이터 수집

Using Python Selenium

02 멜론 [최신곡 - 국내] **Melón**

title	artist	album	date	genre	flac
-------	--------	-------	------	-------	------

like	reply	lyric	lyricist	composer	arranger
------	-------	-------	----------	----------	----------

12개의 노래 변수





02

데이터 수집

Using Python Selenium

03 멜론 [최신곡 - 국내]

MelOn + 지니 [나의 플레이리스트]

genie

title	artist	album	date	genre	flac
-------	--------	-------	------	-------	------

like	reply	lyric	lyricist	composer	arranger
------	-------	-------	----------	----------	----------



Title	Artist
매일 듣는	황치열
빌려줄게	신용재 (포
들리나요	The One (
숨쉬는 모	범키
봄날의 소	예성 (YESU

Cnt
81
71
63
62
56

12개의 노래 변수

곡 정보

감상 횟수



03 데이터 정제





03

데이터 정제

전처리 in R

01 중복 노래 Unique

한잔할까 문명진	한잔할까	2017.11.18R&B / Soul	651명	5개	한잔할까 !문명진,Grc문명진,GrcGroovin
한잔할까 (문명진	한잔할까	2017.11.18R&B / Soul	2107명	19개	한잔할까 !문명진,Grc문명진,GrcGroovin



크롤링의 시간 차로 인해 발생하는 중복 노래 제거





03

데이터 정제

전처리 in R

02 변수처리(날짜)

title	artist	album	date	genre	flac	like	reply	lyric	ist	composer	arranger
On My Mi	박나래	내방을 여	2017.11.22	Dance		좋아요 21	13개	Hey	래	Scoop Deville,Salam	

title	artist	album	genre	flac	like	reply	lyric	Year	Season
On My Mi	박나래	내방을 여	Dance	No_Flac	213	13	Hey Scoop	2017	fall





03

데이터 정제

전처리 in R

02 변수처리(FLAC)

title	artist	album	date	genre	flac	like	reply	lyric	lyricist	composer	arranger
On My Mi	박나래	내방을 여	2017.11.22	Dance		좋아요 21	13개	Hey Scoop	박나래	Scoop Deville,	Salam

title	artist	album	genre	flac	like	reply	lyric	Year	Season
On My Mi	박나래	내방을 여	Dance	No_Flac	213	13	Hey Scoop	2017	fall





03

데이터 정제

전처리 in R

02 변수처리(like / reply)

title	artist	album	date	genre	flac	like	reply	lyric	lyricist	composer	arranger
On My Mi	박나래	내방을 여	2017.11.22	Dance		좋아요 21	13개	Hey Scoop	박나래	Scoop Deville,	Salam

title	artist	album	genre	flac	like	reply	lyric	Year	Season
On My Mi	박나래	내방을 여	Dance	No_Flac	213	13	Hey Scoop	2017	fall





03

데이터 정제

전처리 in R

02 변수처리(작사가/작곡가/편곡가)

title	artist	album	date	genre	flac	like	reply	lyric	lyricist	composer arranger
On My Mi	박나래	내방을 여	2017.11.22	Dance		좋아요 21	13개	Hey Scoop	박나래	Scoop Deville, Salam



title	artist	album	genre	flac	like	reply	lyric	Year	Season
On My Mi	박나래	내방을 여	Dance	No_Flac	213	13	Hey Scoop	2017	fall





03

데이터 정제

전처리 in R

03 가사 공백 노래 삭제

너에게만 (루이, 소유 너에게만 (2017.10.25	Rap / Hip-Flac 16bit 2,851명	8개	너에게만 (루이	루이, Cosm Cosmicboy
그대를 내 Collective Note#5	2017.10.26Ballad Flac 16bit 771명	23개	수줍은 그 유하림	엄태영, 유하림, 엄태영, 유하림
기억 속에 피아노 치#1 소녀의	2017.11.08New Age Flac 16bit 10명	1개		별하 별하
Chemical GATE 9 (GATE1 : P, 2017.10.26	Dance Flac 16bit 171명	5개	A-Yo G9 is BULL\$EYE, 이단옆차기	김동열
품에 (나 노르웨이 노르웨이	2017.10.26New Age Flac 16bit 221명	0개		
나는 새롭새소년 여름기	2017.10.26Rock Flac 16bit 1,046명	5개	눈을 뜬 오 황소윤	황소윤 황소윤, 강토, 문팬시, 김
Mistaken 베일리 (BeLiar	2017.10.27R&B / Soul Flac 16bit 71명	1개	다정하게 베일리 (Be베일리 (Be	주히
미로 김나영 당신이 잠	2017.10.25Drama Flac 16/24 11,607명	34개	눈 감아도 감동is, 서지감동is, 서지감동is, 서재하, 김영성,	
Do it now 태리 (TerrRED&BLUE	2017.10.26Rap / Hip-Flac 16/24 245명	2개	너 하고 싶태리 (Terr태리 (Terr	Cash Note
추억 오늘 추억 - Mo	2017.10.26Folk 83명	0개	구름은 언김지혜	김남일 오늘
한사람 더 히든 한사람	2017.10.26Ballad Flac 16bit 90명	2개	처음 네가 손영채	손영채 손영채, 이기현
Always in Bye Bye B Always in	2017.10.25Rock Flac 16bit 442명	2개	오늘은 특정봉길	Bye Bye B Bye Bye Badman
Laurel 센티멘탈 HISTORY	2017.10.25Electronic Flac 16bit 230명	0개		
그냥 두기 DMEANO 그냥 두기	2017.10.24R&B / Soul Flac 16/24 1,330명	4개	어떻게 지 DMEANO DMEANO DMEANO	(디미너), C
흘러가요 마이큐 흘러가요	2017.10.24R&B / Soul Flac 16bit 891명	4개	파란 하늘 마이큐	마이큐 마이큐
연민 이지연 현이와 덕	2017.10.25Ballad Flac 16/24 46명	1개	사랑은 영장덕	장덕 박강영





03

데이터 정제

전처리 in R

04 장르처리

Animation , Game



Animation/Game

Crossover, Musical



Crossover/Musical

Drama , Korean Move



Drama/Korean Movie

Blues , Jazz



Blues/Jazz





03

데이터 정제

전처리 in R

04 장르처리

[전]

[1]	Adult Contemporary	Animation	Ballad	Blues
[5]	Crossover	Dance	Drama	Electronica
[9]	Electronica, Rock	Folk	Game	Jazz
[13]	Korean Movie	Korean Traditional	Musical	New Age
[17]	Pop	R&B / Soul	Rap / Hip-hop	Rock
[21]	Vocal/Choral	가톨릭음악	국내CCM	기타
[25]	동요	워십	창작동요	

27 Levels: Adult Contemporary Animation Ballad Blues Crossover Dance ... 창작동요

[후]

[1]	Ballad	Folk	R&B / Soul
[4]	Drama / Korean Movie	Rap / Hip-hop	Blues / Jazz / New Age
[7]	Dance	Electronica / Rock	Crossover / Musical
[10]	Animation / Game	Adult Contemporary	

11 Levels: Adult Contemporary Animation / Game ... Rap / Hip-hop





03

데이터 정제

전처리 in R

04 장르처리

[전]

[1]	Adult Contemporary Animation	Ballad	Blues
[5]	Crossover	Dance	Drama
[9]	Electronica, Rock	Folk	Game
[13]	Korean Movie	Korean Traditional Music	New Age
[17]	Vocal/Choral	가톨릭음악	국내CCM
[21]	동요	워십	창작동요
[25]	창작동요		

27 Levels: Adult Contemporary Animation Ballad Blues Crossover Dance ... 창작동요

[후]

[1]	Ballad	Folk	R&B / Soul
[4]	Drama / Korean Movie	Rap / Hip-hop	Blues / Jazz / New Age
[10]	Animation / Game	Adult Contemporary	

11 Levels: Adult Contemporary Animation / Game ... Rap / Hip-hop





03

데이터 정제

전처리 in Python

01 Twitter()를 이용한 Tokenize

*단, Noun / Verb / Adjective 만 사용

02 stop words 제거

03 단어 50개 미만 노래 삭제





03

데이터 정제

전처리 in Python

01 Twitter()를 이용한 Tokenize

02 stop words 제거

03 단어 50개 미만 노래 삭제

- 구글링을 통한 stop words
- 직접 찾기를 통해 수사/대명사/전치사
- 총 808개





03

데이터 정제

전처리 in Python

01 Twitter()를 이용한 Tokenize

02 stop words 제거

03 단어 50개 미만 노래 삭제

총 34032 곡



04 데이터 분석





04

데이터 분석

분석 절차

Word2Vec

가사 유사도



Gower
Similarity

변수들의 유사도



최종 랭킹화





04

데이터 분석

Word2Vec

“You shall know a word
by the company it keeps”

-J.R Firth 1957

단어의 주변을 보면 그 단어를 알 수 있다.





04

데이터 분석

Word2Vec

“유동현 교수님은 원빈 같다.”

“인하대 원빈은 유동현 교수님!”

“겨울에 귤을 까먹는 건 제 맛!”

“후문에서 겨울마다 귤을 판다.”

“겨울의 비타민씨, 귤”





04

데이터 분석

Word2Vec

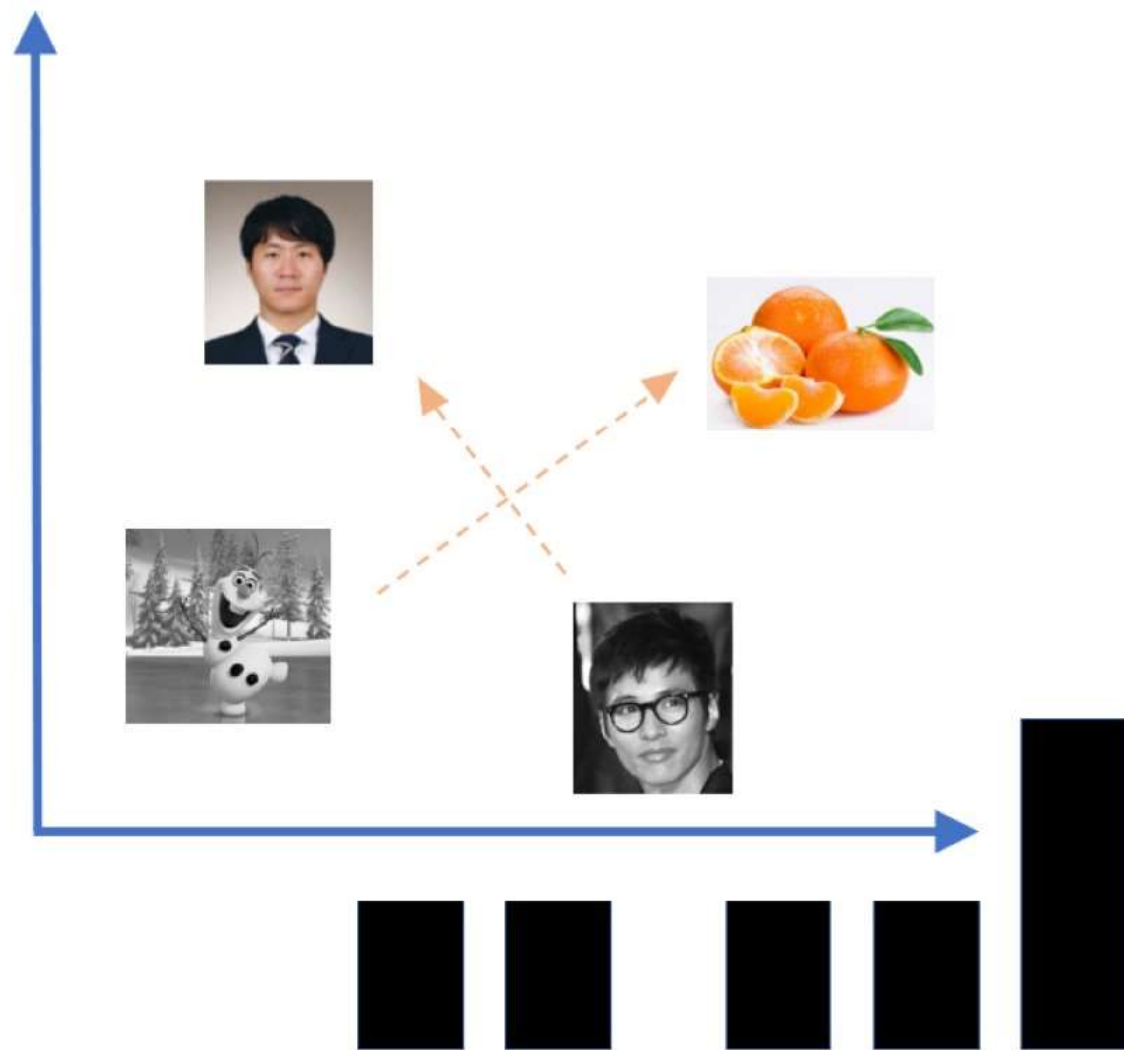
“유동현 교수님은 원빈 같다.”

“인하대 원빈은 유동현 교수님!”

“겨울에 귤을 까먹는 건 제 맛!”

“후문에서 겨울마다 귤을 판다.”

“겨울의 비타민씨, 귤”





04

데이터 분석

Word2Vec

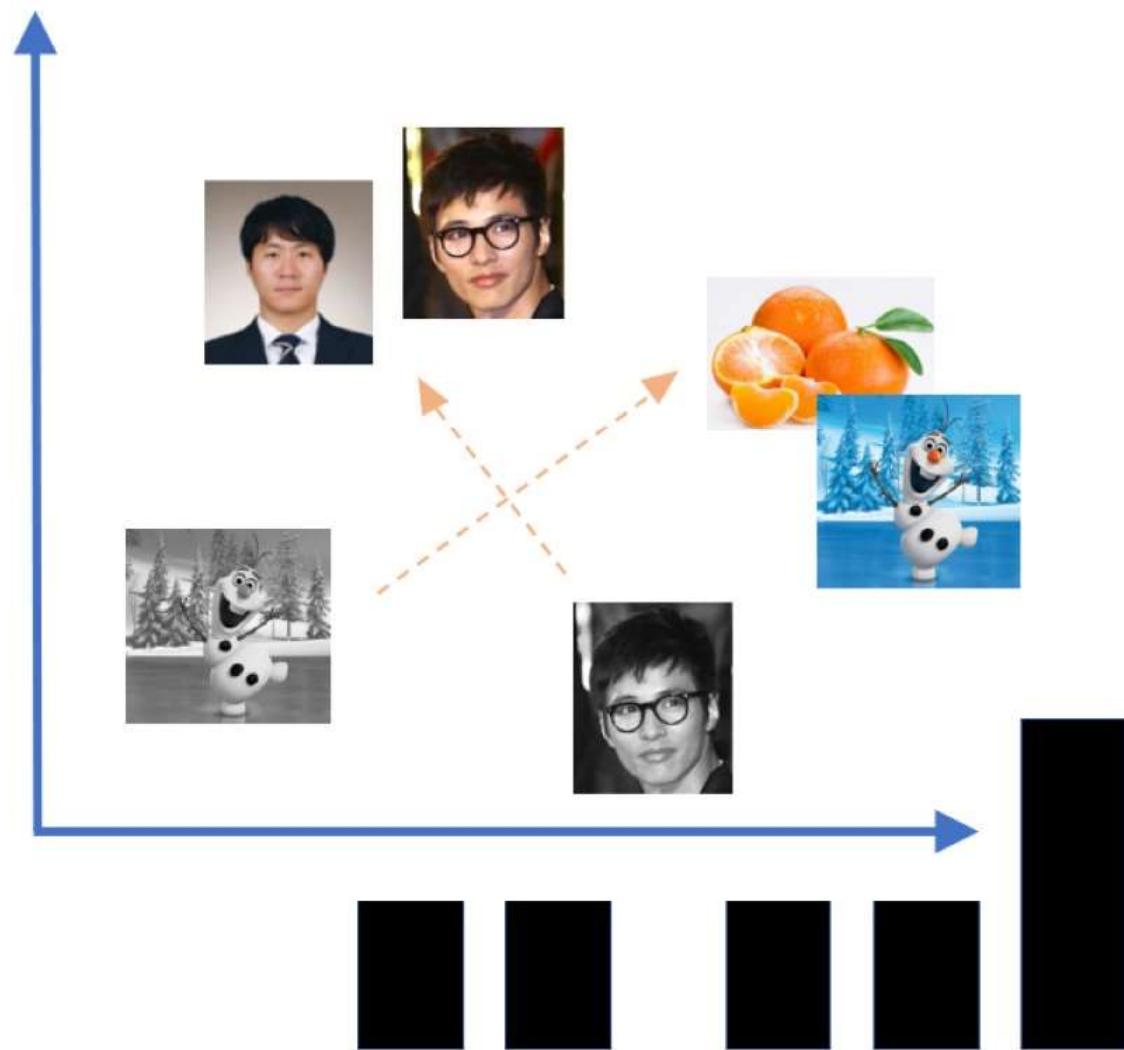
“유동현 교수님은 원빈 같다.”

“인하대 원빈은 유동현 교수님!”

“겨울에 귤을 까먹는 건 제 맛!”

“후문에서 겨울마다 귤을 판다.”

“겨울의 비타민씨, 귤”





04

데이터 분석

Word2Vec

“유동현 교수님은 원빈 같다.”

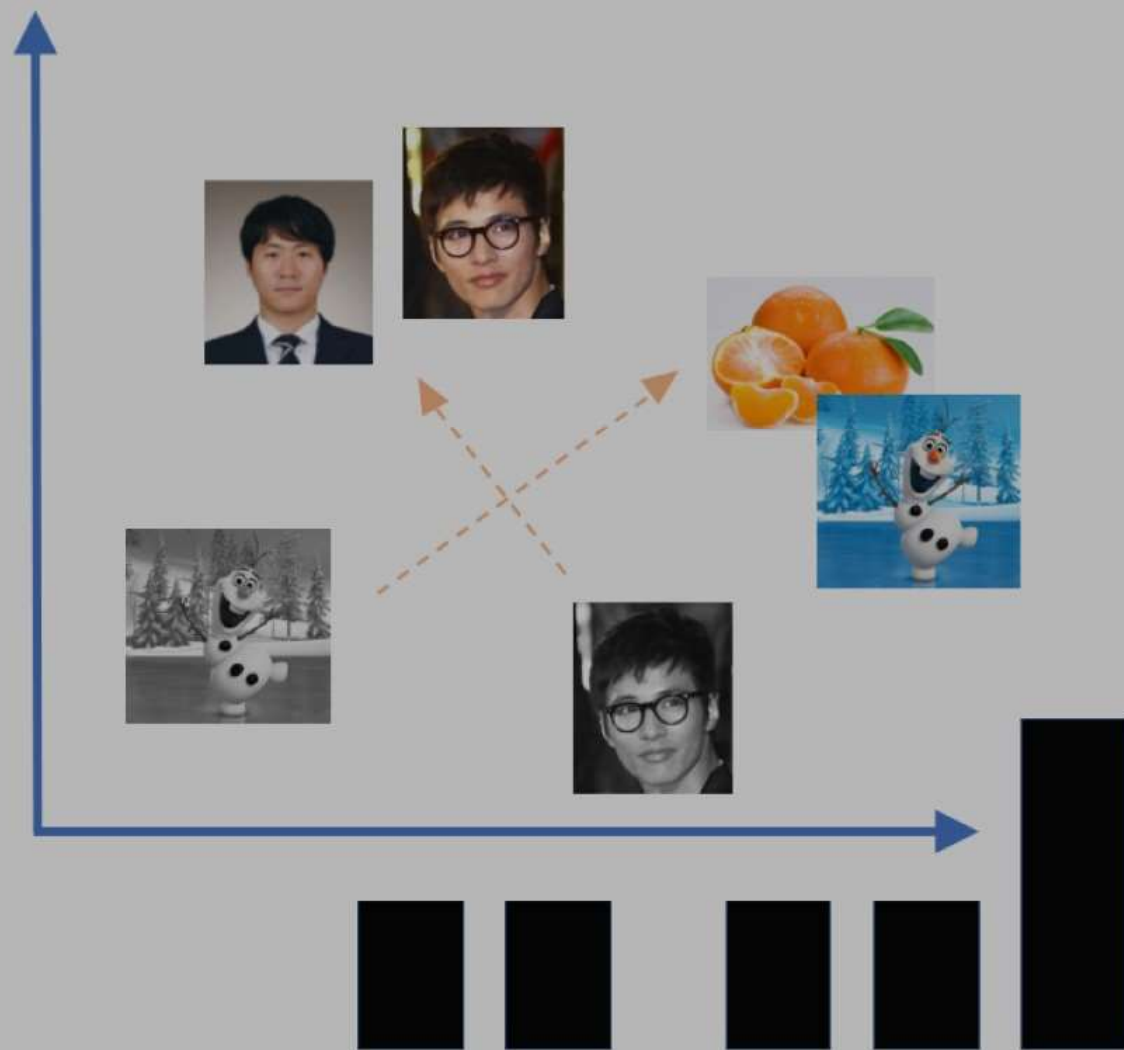
“인하대 원빈은 유동현 교수님!”

“겨울에 귤을 까먹는 건 제 맛!”

“후문에서 겨울마다 귤을 판다.”

Input: 문장

Output: 단어, 단어의 좌표





04

데이터 분석

Word2Vec

[Input]

Twitter()로 형태소 분리된 가사

ex [[...눈/Noun, 내리는/Verb ...],
[...]]



Word2Vec



[Output]

유사도

ID	유사도
1811	0.8470
46	0.1867

10 곡

지니 플레이리스트 290곡 중
최소 6회 이상 재생된 곡 = 80곡

model.n_similarity





04

데이터 분석

Gower Similarity

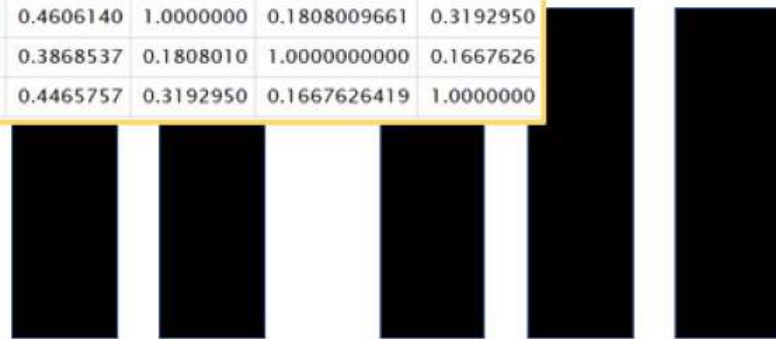
혼합형 변수의 유사도를 구해준다

Binomial
Nominal
Ordinal
Continuous

[플레이리스트 곡]

[Word2Vec 결과
유사도 높은 10곡]

	1	2	3	4	5	6	7	8	9	10	11
1	1.0000000	0.6633230	0.496193359	0.496239427	0.496621772	0.4947399303	0.6623349	0.4520077	0.3247271	0.1721946892	0.8279013
2	0.6633230	1.0000000	0.498208739	0.664921474	0.499965449	0.4967553105	0.8310169	0.4499923	0.4893783	0.1701793090	0.6632500
3	0.4961934	0.4982087	1.000000000	0.499953932	0.498174188	0.6652132382	0.4994748	0.2815344	0.4875871	0.0017213813	0.4983746
4	0.4962394	0.6649215	0.499953932	1.000000000	0.664886923	0.4985005035	0.6660954	0.2815805	0.4876331	0.0017674494	0.4983285
5	0.4966218	0.4999654	0.498174188	0.664886923	1.000000000	0.4967207594	0.4976490	0.2833602	0.3227462	0.0035471934	0.4965488
6	0.4947399	0.4967553	0.665213238	0.498500503	0.496720759	1.0000000000	0.4990717	0.2800810	0.4861337	0.0002679528	0.4996361
7	0.6623349	0.8310169	0.499474834	0.666095432	0.497649022	0.4990717379	1.0000000	0.4476759	0.4870619	0.1678628816	0.6655664
8	0.4520077	0.4499923	0.281534397	0.281580465	0.283360209	0.2800809682	0.4476759	1.0000000	0.4606140	0.3868536513	0.4465757
9	0.3247271	0.4893783	0.487587082	0.487633150	0.322746227	0.4861336534	0.4870619	0.4606140	1.0000000	0.1808009661	0.3192950
10	0.1721947	0.1701793	0.001721381	0.001767449	0.003547193	0.0002679528	0.1678629	0.3868537	0.1808010	1.0000000000	0.1667626
11	0.8279013	0.6632500	0.498374594	0.498328526	0.496548782	0.4996360720	0.6655664	0.4465757	0.3192950	0.1667626419	1.0000000





04

데이터 분석

Ranking

(Lyric similarity x Gower similarity
x 감상횟수의 비율)

의 크기순으로 10곡 선택

총 $80 \times 10 = 800$ 곡 중,
겹치는 노래를 제거해 759곡 이용

ID	Lyric similarity	Gower similarity	감상 횟수 비율
32156	0.9174	0.6920	0.0046
⋮			
21065	0.9341	0.5973	0.0107
30415	0.8787	0.8362	0.0092



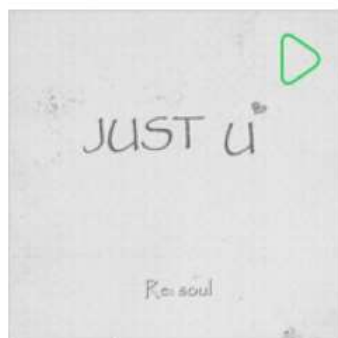
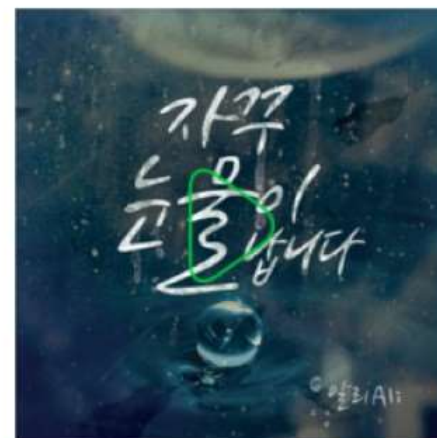
05 분석 결과





05

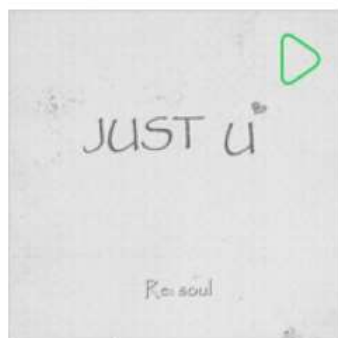
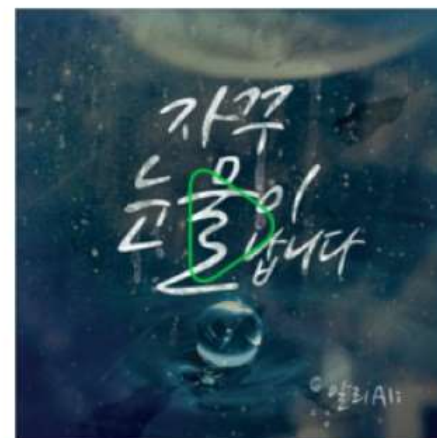
분석 결과





05

분석 결과



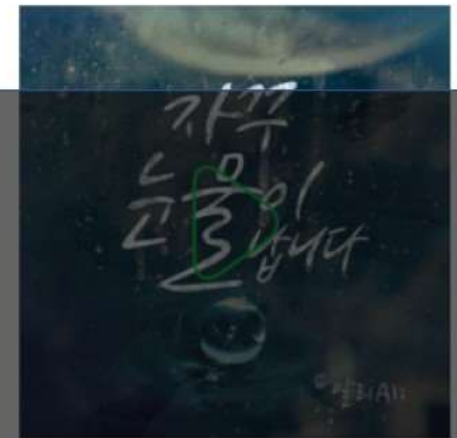
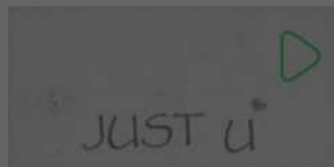


05

분석 결과



-모두 처음 들은 노래
-인지도가 높지 않은 가수
-선호하는 장르의 반영



06 한계 및 제언





06

제언

01 작사가/작곡가/편곡가 변수 고려

	필승불패	배새롬	가을캣	아메리카노
필승불패,배새롬	○	○	X	X
가을캣	X	X	○	X
필승불패,가을캣	○	X	○	X
필승불패,아메리카노	○	X	X	○
필승불패,아메리카노,가을캣	○	X	○	○
필승불패	○	X	X	X
필승불패,가을캣	○	X	○	X





06

제언

01 작사가/작곡가/편곡가 변수 고려

	필승불패	배새롬	가을캣	아메리카노
필승불패, 배새롬	O	O	X	X
가을캣	X	X	O	X
필승불패, 가을캣	O	X	O	X
배새롬, 가을캣	X	X	O	X
필승불패, 배새롬, 가을캣	O	X	X	X
필승불패, 가을캣	O	X	O	X

dummy화 작업을 하기에
지나치게 차원이 커짐



NULL값의 빈도가 높음





06

제언

02 불용어 사전

- Twitter()를 이용한 tokenize가 불명확

03 Gower Similarity

- 범주형 + 연속형의 유사성의 정도





06

제언

04 모델의 성능 판단 불가



[김준식 - 늦은 이별은 건디다]

VS



[H-CODE - 그 날]



Q & A



Thank you

[들려줄게, 너의 노래] 인하대학교 통계학과 빅데이터 자료분석 발표

