

Analysis of Korean Music Lyrics Using Text Mining

2012021037 유영지

2013011075 홍재열

2014021506 송봉근

2015020862 장성원

KOREA

We think the Unthinkable



냉철한 지성이 뜨거운 열정과 손잡는 곳, 세계무대와 통하는 민족지성의 심장
UNIVERSITY

Contents

1. 서론

2. 본론

3. 결론

1. 서론

- 음악이 사회와 문화 속에서 가지는 기능은 음악 그 자체보다는 사람에게 주는 영향력에 있음 (Portnoy, 1963)
- 노래는 가사를 통하여 메시지(개인의 삶 및 사회문화적 가치)를 전달하는 의사소통의 수단



“난 이래서 음악이 좋아.

지극히 따분한 일상의 순간까지도 의미를 갖게 되잖아.

이런 평범함도 어느 순간 갑자기 진주처럼 아름답게 빛나거든.

그게 바로 음악이야”

- 비긴어게인 中



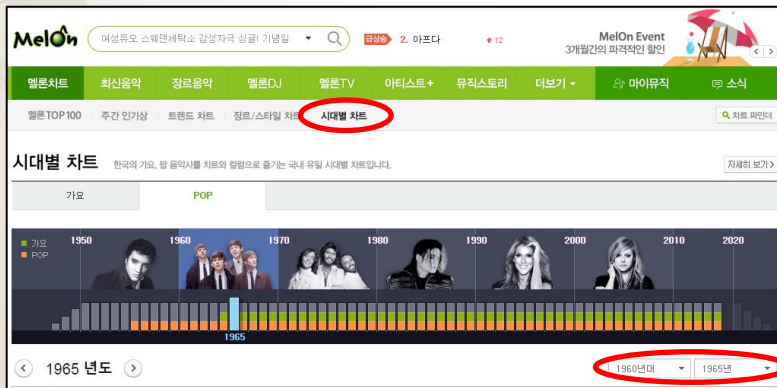
1. 서론

• 목 적

- 한국 대중가요의 가사를 분석하여 시대별 가사의 변화 동향 파악(Clustering)
- 대중 가요에 많이 사용되는 키워드를 주제별로 분석(Topic Modeling)

• 데이터

- 데이터 수집 : 멜론(www.melon.com)의 시대별 차트를 기반으로 시대별 인기곡 선정
- 년도별(1964'~2010') 20개 가사 추출 (1019개)
- 크롤링을 통해 웹사이트에서 가사 추출



년도	장르	제목	가사
1964	Adult Contemporary	내일 또 만나요	하루의 일을 끝내고 돌아가는 거리에 사람의 물결
1964	Adult Contemporary	눈물의 연평도	조기를 담백잡아 기록을 올리고 온다던 그배는 어
1964	Adult Contemporary	님이라 부르리라	임이라 부르리라 당신이라고 부르리라 사랑을 하
1964	Adult Contemporary	동백아가씨	헤일 수 없이 수많은 밤을 내 가슴 도려내는 아픔
1964	Adult Contemporary	떠날 때는 말없이	그 날 밤 그 자리에 돌아서 만났을 때 똑같은 그
1964	Adult Contemporary	마도로스 부기	항구의 일번지 부기우리 일번지 그라스를 채워다
1964	Adult Contemporary	맨발의 청춘	눈물도 한숨도 나 혼자 씹어 삼키며 밤 거리의 뒷
1964	Adult Contemporary	밀짚모자 목장아가씨	시원한 밀짚모자 포플라 그늘에 양떼를 물고가는
1964	Adult Contemporary	보고싶은 얼굴	눈을 감고 걸어도 눈을 뜨고 걸어도 보이는 것은
1964	Adult Contemporary	빗속의 여인 (1964년작)	잊지 못할 빗속의 여인 지금은 어디 있나? 노오
1964	Adult Contemporary	빨간구두 아가씨	솔솔솔 오솔길에 빨간구두 아가씨 똑똑똑 구두소
1964	Adult Contemporary	소쩍새 우는 마을	소쩍새 울고간 뒤에 나풀나풀 나비가 춤추며
1964	Adult Contemporary	아메리칸 마도로스	무역선 오고가는 부산 항구 제2부두 죄많은 마도
1964	Adult Contemporary	아빠 안녕	비가 나리네 비가 나리네 돌아선 이가슴에 그님

2. 본론

데이터전처리

- POS Tagging, Stop Words List를 바탕으로 의미 없는 단어 제거
- 한글 / 영어를 구분하여 데이터 전처리 실시
- 한글 데이터 전처리의 문제점
 - 대부분의 자연어 처리 패키지가 영어에 초점이 맞춰짐 한글은 stemming, lemmatization을 바로 해주는 R-패키지가 없음
 - Tm 패키지의 TermDocumentMatrix 함수를 사용시 "가슴 사랑", "그리고 사랑은" 등 두 어절 이상의 단어가 하나의 Term으로 구성되는 문제점 발생

2. 본론

데이터전처리

➤ 전처리 방법 (한글)

- POS tagging을 이용한 노래가사 품사(22가지) 분류

* 패키지 : KoNLP 패키지_ SimplePOS22 함수

미려다	미쳐	내가
"미르/PA+머다/EC"	"미치/PV+머/EC"	"나/NP+가/JC"
그런	내가	너
"그러/PV+L /ET"	"나/NP+가/JC"	"너/NP"
내가	독한	나로
"나/NP+가/JC"	"독하/PA+L /ET"	"나/NP+로/JC"
널	앓은	인형에다
"널/NC"	"앓/PV+은/ET"	"인형/NC+에/JC+미/JP+다/EF"
걸어	내가	그녀와
"걸/PV+머/EC"	"나/NP+가/JC"	"그녀/NP+와/JC"
고	every	night
"고/MM"	"every/F"	"night/F"
be	with	you
"be/F"	"with/F"	"you/F"
love	her	do
"love/F"	"her/F"	"do/F"
her	매일	나의
"her/F"	"매일/MA"	"나/NP+의/JC"

* 결과 : 문자열 처리 → 불필요 문자 제거 → 가사별 벡터 재구성

2. 본론

➤ 전처리 방법 (한글)

- 분석 목적에 따라 필요한 품사 선택
 - * 토픽 모델링: 보통명사(NC), 동사(PV), 형용사(PA), 외국어(F)
 - * 시대별 가사: 대명사(NP), 종결어미(EF), 보통명사, 동사, 형용사, 부사
- Stopword 제거
 - * 기존에 존재하는 한글 stopwords 리스트는 분석에 필요 단어들이 다수 존재
→ 가사의 추세를 살펴보고 적절한 stopwords 선택
 - * Term-DF(Document Frequency)를 통해 불필요 단어를 stopwords로 지정

종결어미	부사	형용사	대명사	동사
-다, -ㄴ다	그리고, 그러나 등	그러다, 안하다 등	나, 너, 그것 등	그렇다 등

➤ 전처리 방법 (영어)

- 형태소 분석이 이루어진 한글과 달리 영어는 love, loves 등이 공존
→ lemmatization → tm패키지의 stopwords("english")
→ DF를 바탕으로 만든 stopwords를 이용해 불필요한 단어 제거

➤ Term-Document Matrix 구성 <Term:8569개 / Document:1018개>

2. 본론

➤ Term-Doc Matrix의 Term부분 비교

"가슴 눈길"	"가을 가을 낙엽"	"걸어가는걸"
"거짓말 "	"brain"	"걸 갈대"
"가재 모습"	"faster"	"갑 시집"
"거짓말 바보"	"bright"	"catch"
"간대네 저녁 노을"	"간곳이 물어물어"	"고치곤해"
"거문목화"	"found"	"가슴 기도"
"가슴 기도"	"가슴 마음"	"강 산을끼고"
"finally"	"견딜수있을거야 "	"개 계란"
"miss"	"가슴 사랑한거야 눈 "	"걸 공간"
"가슴 사람"	"강 등불"	"거짓말 가슴 "
"갈바랍 외로움"	"갯바위 당신"	"거짓말 외톨이 "
"갯마을 "	"가슴 새 "	"개옹벌레"
"가슴 떼어논채"	"겨울가고 마음"	"걸 그댄"
"가슴앓고 오늘"	"beyonce"	"고무신\n"
"갈을게 못준게 "	"가슴 슬프긴 당신"	"pump"
"가슴 꿈 만날길 바람"	"견딜수있을꺼야 "	"가슴하나로"
"가슴 기운"	"back"	"가신님이 가지"
"갈바랍 그대"	"건가봐 법 세상"	"걸 떠난님이"
"걸 바보"	"boy"	"것들을 "
"계절가네 어이허나"	"bring"	"가슴 차 두볼에 "
"가을 사람"	"가슴 아아아 너를안고"	"가을 석류"
"갈 갈 "	"가슴 죄책감 외면"	"거짓말 얘기"
"것들을 거짓"	"가슴 한가운데"	"가위 "
"allright"	"갈대발 "	"갈곳을 "
"at1"	"개 일곱색깔"	"거라곤 "
"light"	"cuatro"	"경포대"
"가슴 없을꺼야"	"eyes"	"고운정만"
"가슴 자연"	"give"	"beutiful"
"가식 "	"playboy"	"갈길만가네 길"

tm 패키지 사용



"oh"	"lolly"	"put"
"baby"	"gee"	"사랑"
"hot"	"사랑"	"여자"
"love"	"아름답다"	"woo"
"girl"	"여인"	"모르다"
"ah"	"모습"	"살다"
"whoo"	"사람"	"떠나다"
"heart"	"있다"	"hand"
"body"	"지나다"	"세상"
"give"	"떠나다"	"알다"
"bad"	"슬프다"	"좋다"
"issue"	"lollipop"	"아프다"
"go"	"마음"	"마음"
"nee"	"만나다"	"가슴"
"철없다"	"생각"	"많다"
"let"	"바람"	"사랑해"
"know"	"모르다"	"사람"
"hallo"	"고향"	"sorry"
"get"	"기다리다"	"웃다"
"name"	"알다"	"있다"
"make"	"찾다"	"늦다"
"어쩌다"	"볼다"	"흐르다"
"want"	"울다"	"울다"
"sum"	"추억"	"노래"
"아리송해"	"가슴"	"걸다"
"dj"	"그립다"	"보내다"
"back"	"이젠"	"외톨이"
"미치다"	"dat"	"슬프다"
"예쁘다"	"피다"	"만나다"
"bye"	"오늘"	"싫다"

패키지 사용 X

2. 본론

시대별 가사 분석

➤ 방법

- Matrix 재구성

	song1	song2	song3	Song 1018
사랑	6	2	8	10
자유	2	5	0	9
누군지	3	7	6	3
...
...
하루	5	8	0	0	0	0

<Binary Transform>

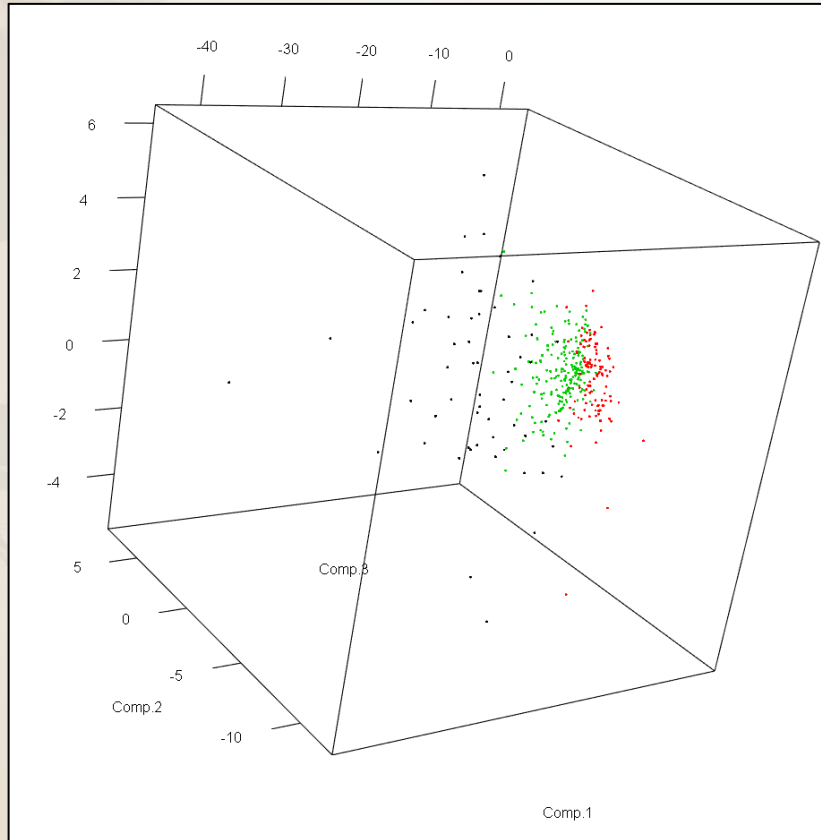
	Song1	song2	song3	Song 1018
사랑	1	1	1	1
자유	1	1	0	1
누군지	1	1	1	1
...
...
하루	1	1	0	0	0	0

	Year 1964	Year 1965	Year 1966	Year 2014
사랑	20	9	5	7
자유	4	6	0	9
누군지	6	7	8	0
...
...
하루	8	10	14	7	9	0

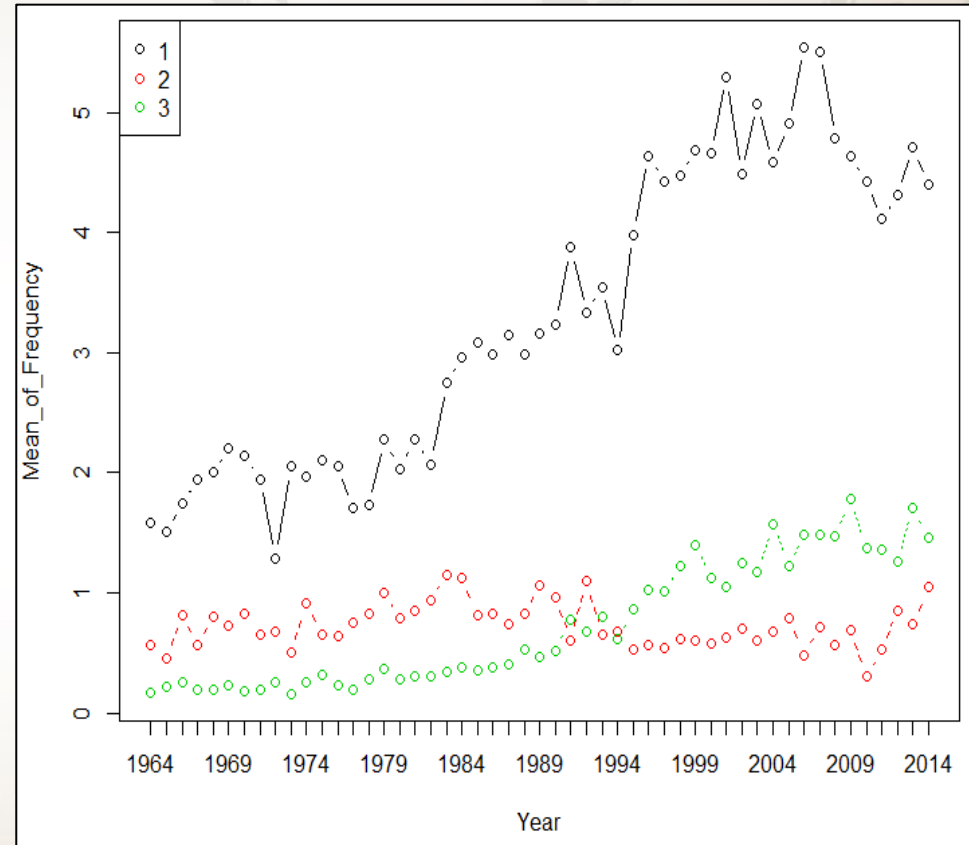
- * 단순빈도수: 특정 노래에서 많이 사용된 단어는 그 시대를 대표하는 단어가 아님
- DF가 큰 단어들로 Clustering에 사용할 Keyword 선정(360개 단어)
- Hierarchical Clustering 사용
 - 유사도 척도 : Jaccard Coefficient / 군집 방법 : Ward's method
 - Sparse한 데이터의 특성으로 인해서 Euclid distance 대신 Jaccard 사용
 - Single, complete, average 모두 군집이 한 군집에 많이 몰려서 Ward 방법 사용

2. 본론

시대별 가사 Clustering 결과



<PCA 3D Plot>



<Frequency of Keywords in each Clusters over the Time>

Y축: (각 군집의 년도별 단어 빈도의 합)/(군집크기)

2. 본론

시대별 가사 Clustering 결과

[[1]]												
[1]	"하루"	"사람"	"하늘"	"마음"	"눈물"	"사랑"	"가슴"	"오늘"	"모습"	"얼굴"	"혼자"	"세상"
[13]	"행복"	"추억"	"생각"	"모두"	"지금"	"이제"	"정말"	"이젠"	"그대"	"우리"	"누구"	"돌다"
[25]	"만나다"	"가다"	"바라보다"	"부르다"	"올다"	"있다"	"기다리다"	"떠나다"	"걸다"	"알다"	"지나다"	"흐르다"
[37]	"찾다"	"잡다"	"남다"	"모르다"	"밀다"	"위하다"	"보내다"	"살다"	"들다"	"웃다"	"힘들다"	"아프다"
[49]	"슬프다"	"멀다"	"좋다"	"많다"	"다시"	"아직"	"자꾸"					
[[2]]												
[1]	"거리"	"가로등"	"바람"	"가슴속"	"그리움"	"꽃잎"	"자연"	"달빛"	"태양"	"여인"	"겨울"	"소리"
[13]	"파도"	"구름"	"노래"	"미련"	"기대"	"시절"	"그림자"	"고향"	"운명"	"옛날"	"바다"	"낙엽"
[25]	"외로움"	"한마디"	"세월"	"내일"	"이야기"	"고개"	"마음속"	"참가"	"아래"	"그날"	"그때"	"햇살"
[37]	"불빛"	"어깨"	"영원"	"어둠"	"날들"	"뒷모습"	"가득"	"당신"	"우리들"	"무엇"	"스치다"	"일다"
[49]	"젖다"	"달래다"	"피다"	"감다"	"쓰다"	"울리다"	"떨어지다"	"안기다"	"떠오르다"	"들리다"	"앉다"	"넘치다"
[61]	"그리다"	"사라지다"	"불다"	"마르다"	"헤매다"	"잠들다"	"태우다"	"쌓이다"	"감추다"	"내리다"	"맺히다"	"꺼지다"
[73]	"속삭이다"	"날다"	"이루다"	"잊히다"	"설레이다"	"그립다"	"수많다"	"빨강다"	"졸겁다"	"비다"	"외롭다"	"푸르다"
[85]	"꼭다"	"어둡다"	"서럽다"	"어리다"	"깊다"	"노랑다"	"뜨겁다"	"아름답다"	"하얗다"	"말없이"	"영원히"	"멀리"
[97]	"홀로"	"한없이"	"조용히"	"더욱"	"구나"	"리라"	"오리"	"는가"	"어마지"			
[[3]]												
[1]	"잘못"	"자리"	"순간"	"상처"	"처음"	"입술"	"웃음"	"이별"	"머리"	"이름"	"어제"	"향기"
[13]	"인생"	"생각해"	"남자"	"바보"	"여자"	"거짓말"	"사이"	"미소"	"서로"	"행복한"	"친구들"	"사랑해"
[25]	"눈빛"	"후회"	"대답"	"애기"	"목소리"	"아침"	"악속"	"내게로"	"그곳"	"오늘밤"	"친구"	"사람들"
[37]	"시작"	"마지막"	"그때"	"그때"	"기분"	"시간"	"기억"	"표정"	"그댄"	"자신"	"소중한"	"이유"
[49]	"이상"	"걱정"	"이해"	"예전"	"미안"	"전화"	"모든것"	"조금씩"	"미안해"	"종일"	"매일"	"심장"
[61]	"계속"	"언제"	"어디"	"그녀"	"아무"	"누군가"	"아무것"	"헤어지다"	"지치다"	"채우다"	"깨다"	"보이다"
[73]	"걸다"	"뒤돌다"	"돌리다"	"보다"	"쏟다"	"사다"	"묻다"	"타다"	"먹다"	"채우다"	"변하다"	"풀다"
[85]	"뛰다"	"지키다"	"따르다"	"흔들리다"	"태어나다"	"바라다"	"넘다"	"자다"	"맞다"	"갖다"	"읽다"	"쉬다"
[97]	"다가오다"	"원하다"	"끝나다"	"참다"	"치다"	"죽다"	"받다"	"남기다"	"떨리다"	"지내다"	"달리다"	"애쓰다"
[109]	"견디다"	"서다"	"전하다"	"미치다"	"흘리다"	"빠지다"	"숨다"	"쓰다"	"꿈꾸다"	"나누다"	"마시다"	"흔들다"
[121]	"열다"	"꾸다"	"숨기다"	"만들다"	"나오다"	"들다"	"비추다"	"취하다"	"마주치다"	"어떡하다"	"멈추다"	"느끼다"
[133]	"가지다"	"향하다"	"집다"	"불잡다"	"피하다"	"달다"	"예쁘다"	"방다"	"다르다"	"크다"	"차갑다"	"늑다"
[145]	"짧다"	"밟다"	"기쁘다"	"싫다"	"까맣다"	"길다"	"작다"	"괜찮다"	"바쁘다"	"눈부시다"	"쉽다"	"두렵다"
[157]	"힘겹다"	"고맙다"	"얼마나"	"아무리"	"차라리"	"그리"	"도대체"	"몰래"	"함께"	"먼저"	"아마"	"아주"
[169]	"우연히"	"어느새"	"그냥"	"항상"	"혹시"	"벌써"	"조금"	"오직"	"잠시"	"제발"	"바로"	"웬지"
[181]	"가끔"	"마치"	"모든게"	"이미"	"절대"	"는거야"	"good"	"think"	"know"	"like"	"say"	"love"
[193]	"want"	"baby"	"yeah"	"get"	"go"	"let"	"oh"	"girl"				

2. 본론

시대별 가사 Word Cloud 결과

<군집 1>



<군집 2>



<군집 3>



2. 본론

시대별 가사 Clustering & Word Cloud 결과 분석

군집1 : 과거부터 지금까지 노래가사에 빈번히 사용되는 단어

- * 사람, 마음, 눈물, 사랑, 가슴, 행복, 추억, 아프다, 슬프다 : **감정**과 관련된 단어가 많음
- 노래 가사는 시대를 막론하고 사람들이 보편적으로 느끼는 감정을 대변하는 경우가 많음
- 최근에 빈도수가 급격히 증가하는 이유는 노래 가사의 소재가 과거보다 획일화되었기 때문이라고 유추

군집2 : 과거에 많이 쓰이다 최근에 덜 쓰이는 단어

- * 바람, 꽃잎, 달빛, 태양, 파도, 구름, 바다, 낙엽, 햇살, 어둠, 푸르다 : **자연 친화적**(비유적)인 단어
- * 가로등, 가슴속, 사연, 미련, 운명, 옛날, 세월, 창가, 스치다, 그리움, 아름답다, 속삭이다 : **감성적**인 단어
- * 옛날, 세월, 여인, 당신 -구나, -리라, -우리, -는가, -어야지 : **옛스러운** 단어와 종결어미
- 과거에 많이 사용되던 단어는 최근보다 소재가 좀 더 다양하다

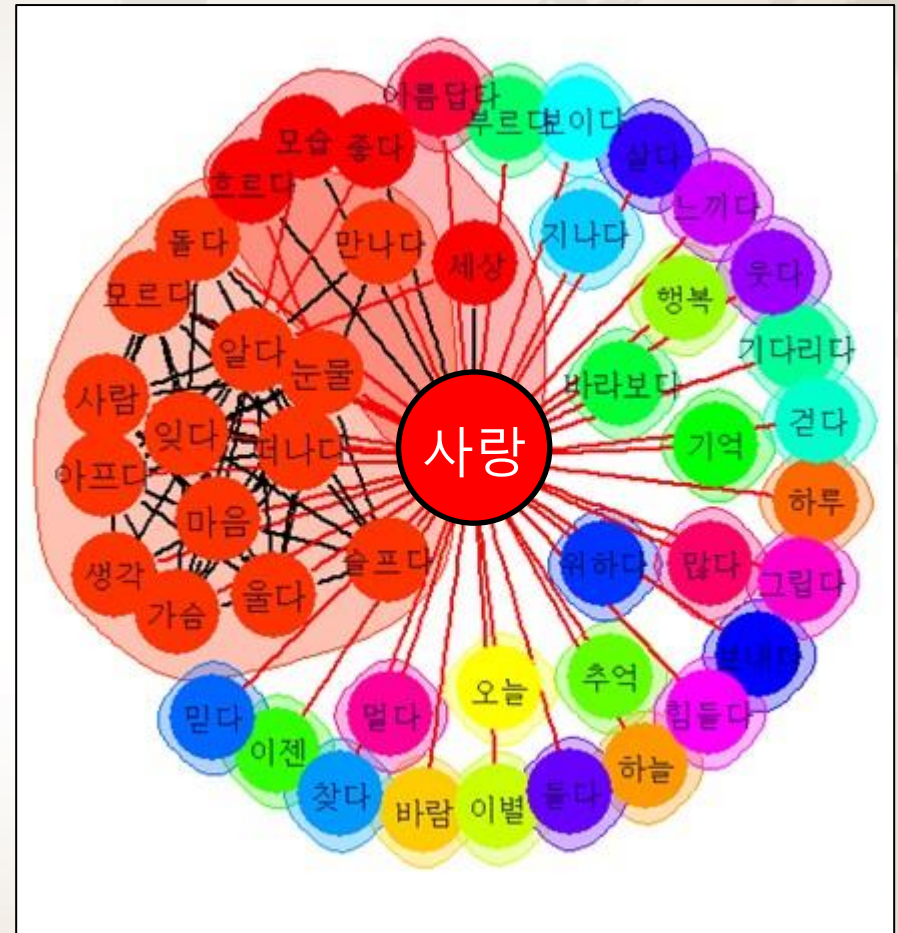
군집3 : 과거에 덜 쓰이다 최근에 많이 쓰이는 단어

- * 순간, 입술, 눈빛, 표정, 오늘밤, 취하다, 변하다, 흔들리다, 다가오다 : **충동적, 외면적**인 단어
- * 바보, 생각해, 사랑해, 미안해, 예쁘다, 싫다 : **직접적인 표현**
- * 이별, 상처, 헤어지다, 지우다 : **이별**에 관한 표현 증가
- * oh, baby, love, think, know, like, want, girl, -는거야, 내게로 : **영어**의 사용과 **현대적** 표현
- 소재가 다양화 된 것은 아니지만 시간이 지남에 따라서 빈도가 증가함

2. 본론

Keyword Frequency Network Plot

- 빈도수가 높은 단어로 Network Plot 형성
전체적인 키워드의 연관성을 파악
- '사랑'이라는 Topic을 중심으로 커뮤니티 형성
(세부적인 Topic에 대한 정보가 부족함)
- Topic별로 시대별 흐름을 파악하기 위해서
Topic Modeling 수행



2. 본론

토픽모델링

➤ 방 법

- Term-Document Matrix → TF-IDF로 구성
(여러 노래에 존재하는 단어들은 Topic이라 볼 수 없는 경우가 많기 때문)
- 대중가요에 많이 사용되는 키워드를 주제별로 분석
(적절한 Topic Modeling 알고리즘(LDA) 사용)
- 토픽 분포에 대한 사전지식이 없기 때문에 알파값과 토픽수를 변경하며 실험
 - Inference Method : Gibbs Sampling
 - Topic 수 : 1, 2, 3, 4, 5, 6, 7, 8, 9, 10
 - ALPHA : 0.01, 0.1, 1, 10, 100

2. 본론

토픽모델링 결과 <Topic - Term>

Topic1 : 자연, 여행에서 나오는 흥
 - 나에게로 떠나는 여행, 두바퀴로가는
 자동차, 산너머 남촌에는 등

Topic2 : 스타일에 관련된 후크송
 - Fire, 유고걸, 강남스타일,
 hot summer, hot issue 등

Topic3 : 클럽댄스 느낌의 후크송
 - lollipop, love, It's raining
 touch my body, loving U 등

Topic4 : 사랑의 아픔, 이별(직접적)
 - 수취인불명, 날울리지마, 기도, 사랑
 하기 때문에, 부디 등

Topic5 : 사랑의 아픔, 이별(간접적)
 - 길가에 앉아서, 달타령, 여인아, 하늘
 못다핀 꽃 한송이 등

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
[1,]	"whoo"	"oh"	"bounce"	"사랑"	"바람"
[2,]	"gee"	"put"	"lollip"	"아프다"	"huh"
[3,]	"heaven"	"hot"	"oh"	"알다"	"마음"
[4,]	"사랑"	"love"	"ah"	"살다"	"내리다"
[5,]	"hallo"	"baby"	"nobody"	"떠나다"	"잊다"
[6,]	"예예예"	"girl"	"baby"	"눈물"	"여인"
[7,]	"사랑해"	"like"	"lone"	"사람"	"하얏다"
[8,]	"아름답다"	"eh"	"dan"	"잊다"	"거리"
[9,]	"좋다"	"thi"	"woo"	"슬프다"	"하늘"
[10,]	"파도"	"cool"	"body"	"이젠"	"피다"
[11,]	"나혼자"	"yeah"	"love"	"모르다"	"불다"
[12,]	"세상"	"철없다"	"bye"	"울다"	"그립다"
[13,]	"날다"	"get"	"heart"	"돌다"	"사람"
[14,]	"마음"	"go"	"lollipop"	"이별"	"모습"
[15,]	"노래"	"give"	"sorry"	"세상"	"걸다"
[16,]	"바다"	"bad"	"time"	"많다"	"비다"
[17,]	"정말"	"hand"	"dance"	"만나다"	"흐르다"
[18,]	"바야야"	"say"	"good"	"uh"	"shake"
[19,]	"불다"	"come"	"want"	"웃다"	"고향"
[20,]	"하늘"	"issue"	"아리송해"	"밀다"	"아름답다"
[21,]	"여자"	"hey"	"life"	"좋다"	"가슴"
[22,]	"워워"	"nee"	"let"	"가슴"	"노래"
[23,]	"즐겁다"	"rock"	"bravo"	"그댈"	"떠나다"
[24,]	"라라라라"	"know"	"댄스"	"마음"	"사람"
[25,]	"내마음"	"back"	"rain"	"기억"	"랄라"
[26,]	"부르다"	"name"	"이젠"	"헤어지다"	"울다"
[27,]	"떠나다"	"feel"	"기다리다"	"위하다"	"찾다"
[28,]	"기다리다"	"make"	"랄랄랄"	"모습"	"외롭다"
[29,]	"사내"	"외톨이"	"알거야"	"기다리다"	"슬프다"
[30,]	"타다"	"sum"	"alright"	"보내다"	"얼굴"

2. 본론

토픽모델링 결과

<Topic – Document>

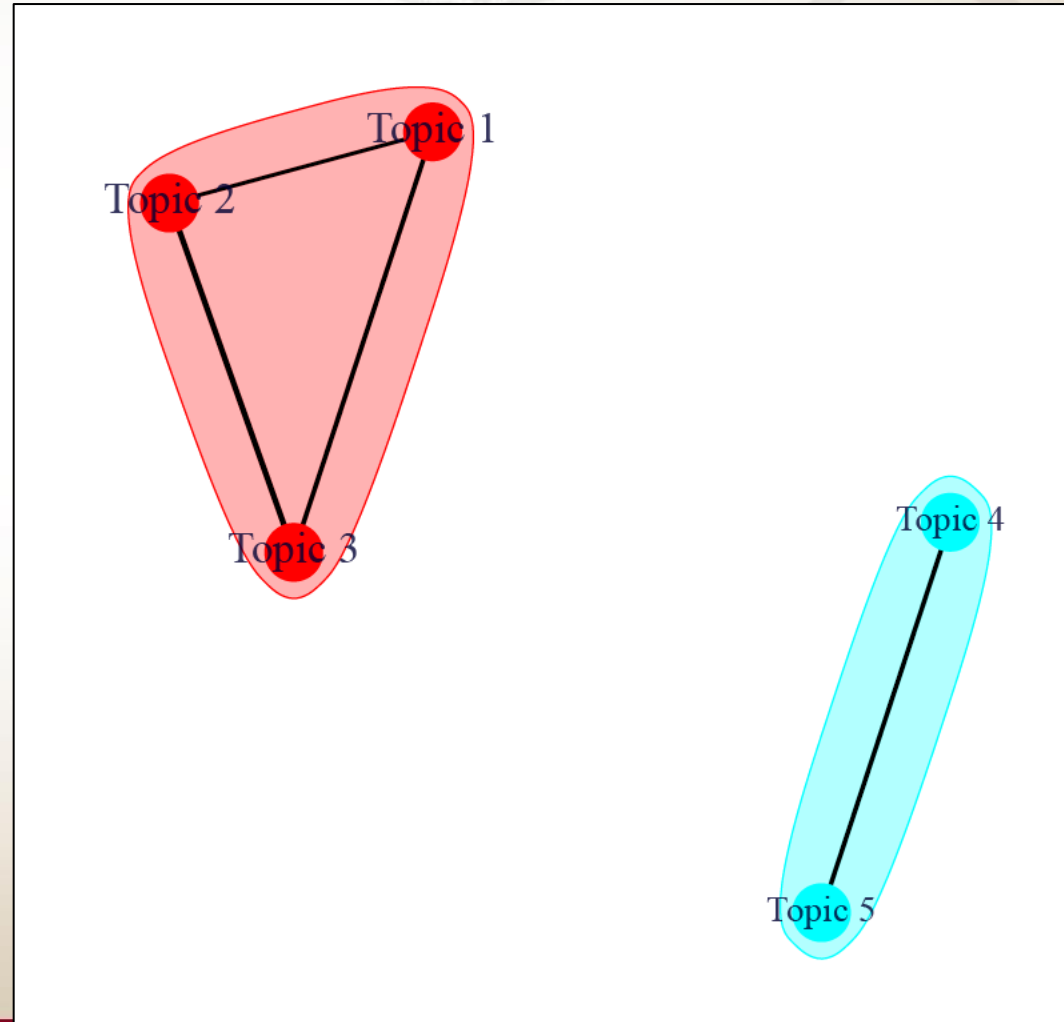
Topic	Posterior	Title	Year	Genre
1 Topic_1	0.990147783251232	나에게로 떠나는 여행	2005	Rock
2 Topic_1	0.989208633093525	Doc와 춤을	1997	Dance
3 Topic_1	0.989100817438692	두 바퀴로 가는 자돌차	1995	Folk
4 Topic_1	0.988919667590028	조용한 여자	1974	Folk
5 Topic_1	0.988235294117647	산너머 남촌에는	1965	Adult Contemporary
6 Topic_1	0.9875	매직 카펫 라이드	2000	Rock
7 Topic_1	0.987460815047022	라라라	1971	Folk
8 Topic_1	0.987301587301587	마포사는 황부자	1965	Adult Contemporary
9 Topic_1	0.986013986013986	내 강산 좋은시구	1973	Adult Contemporary
10 Topic_1	0.985294117647059	희망사항	1990	Ballad
11 Topic_2	0.997262149212868	Hands Up	2011	Dance
12 Topic_2	0.996592844974446	Fire	2009	Dance
13 Topic_2	0.996450754214729	U-Go-Girl	2008	Dance
14 Topic_2	0.995222929936306	So Cool	2011	Dance
15 Topic_2	0.994780793319415	Good Life	2001	Rap / Hip-hop
16 Topic_2	0.993865030674847	강남스타일	2012	Rap / Hip-hop
17 Topic_2	0.993859649122807	Hot Summer	2011	Dance
18 Topic_2	0.993414863593603	Love	2010	Rock
19 Topic_2	0.993413830954994	Hot Issue	2009	Dance
20 Topic_2	0.992402659069326	Liquor Shots	2004	Rap / Hip-hop
21 Topic_3	0.996434937611408	Lollipop	2009	Dance
22 Topic_3	0.995820271682341	쏘리 쏘리	2009	Dance
23 Topic_3	0.995636998254799	FANTASTIC BABY	2012	Dance
24 Topic_3	0.995288574793875	GENTLEMAN	2013	Dance
25 Topic_3	0.994993742177722	Run To You	2000	Rap / Hip-hop
26 Topic_3	0.994285714285714	Loving U	2012	Dance
27 Topic_3	0.99266862170088	Touch My Body	2014	Dance
28 Topic_3	0.992592592592593	It's Raining	2004	Dance
29 Topic_3	0.98970398970399	UGLY	2011	Dance
30 Topic_3	0.9890625	Nobody	2008	Dance
31 Topic_4	0.992088607594937	날개 잃은 천사	1995	Dance
32 Topic_4	0.991341991341991	수취인 불명	2007	Rap / Hip-hop
33 Topic_4	0.989539748953975	칼데일 사랑	1994	Dance
34 Topic_4	0.989189189189189	시간을 찾아서	2004	Rap / Hip-hop
35 Topic_4	0.988980716253444	하늘 나라 우리님 1	1985	Rock
36 Topic_4	0.988472622478386	날 울리지마	1991	Ballad
37 Topic_4	0.987577639751553	기도	2000	Drama
38 Topic_4	0.987179487179487	마는사람 얘기	2013	Rap / Hip-hop
39 Topic_4	0.985815602836879	사랑하기 때문에	2006	Ballad
40 Topic_4	0.98567335243553	부디	1995	Ballad
41 Topic_5	0.989473684210526	나 이런사람이야	2010	Rap / Hip-hop
42 Topic_5	0.989333333333333	독도는 우리땅	1982	Folk
43 Topic_5	0.987460815047022	날씬한 아가씨끼리	1966	Adult Contemporary
44 Topic_5	0.987437185929648	길가에 앉아서	1975	Folk
45 Topic_5	0.985347985347985	달타령	1972	Adult Contemporary
46 Topic_5	0.984251968503937	그녀의 웃음소리뿐	1987	Ballad
47 Topic_5	0.983739837398374	그댄 봄비를 무척 좋아하나요	1985	Folk
48 Topic_5	0.983498349834984	여인아	1982	Adult Contemporary
49 Topic_5	0.983193277310924	하늘	1973	Folk
50 Topic_5	0.983122362869198	못다핀 꽃 한 송이	1984	Rock

2. 본론

토픽모델링 결과 <Topic Network Plot>

토픽간의 연관성을 포함한 토픽모델링 결과 검증

- Community 1
 - Topic 1, Topic 2, Topic 3
- Community 2
 - Topic 4, Topic 5



2. 본론

토픽모델링 결과 <각 토픽의 시대별 흐름 분석>

Hot Topic

클럽느낌의 후크송(Topic3)

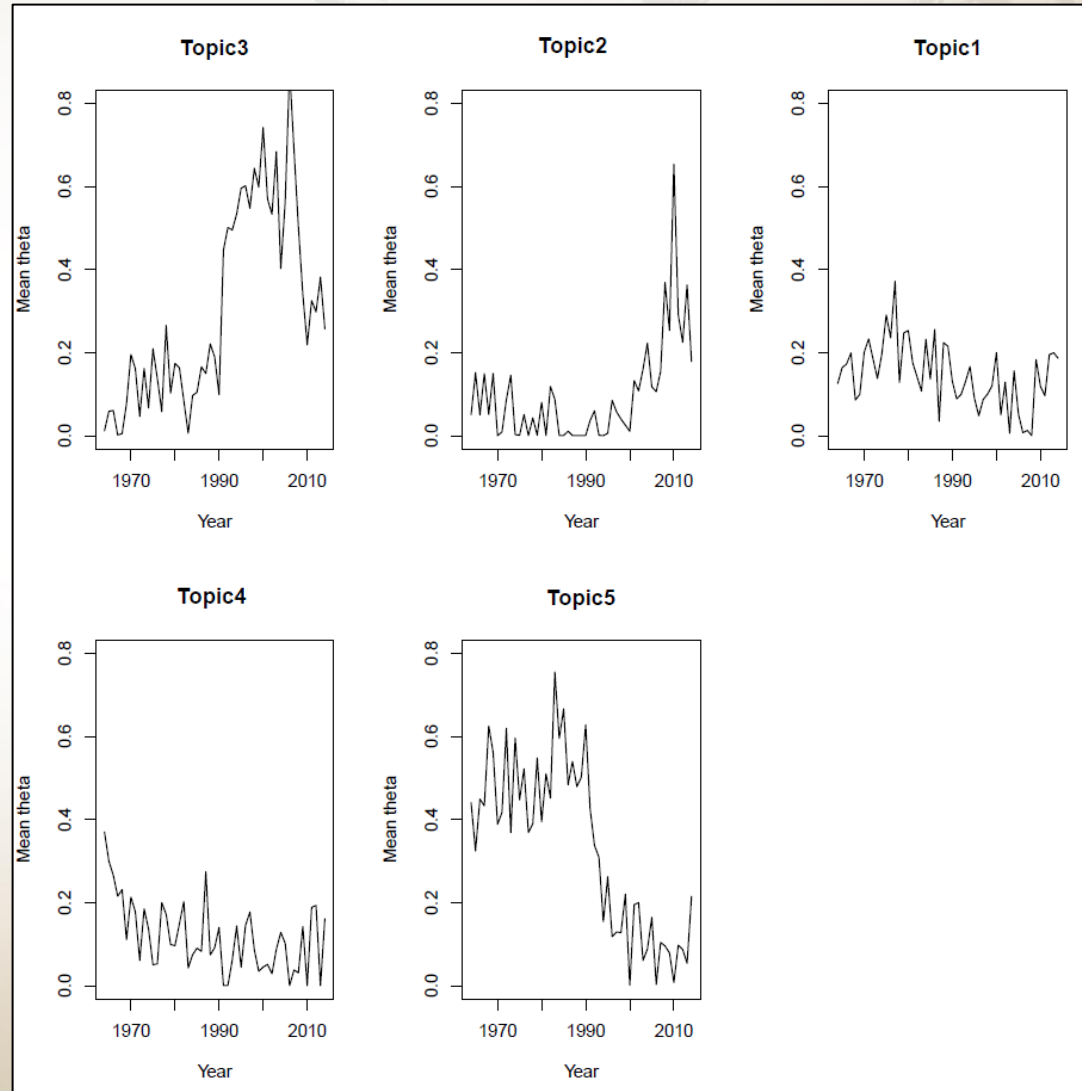
스타일에 관련된 후크송(Topic2)

자연, 여행에서 나오는 흥(Topic1)

직접적인 사랑의 아픔, 이별 (Topic4)

간접적인 사랑의 아픔, 이별 (Topic5)

Cold Topic



3. 결 론

1. 시대별 가사의 변화 동향 파악

- > 새로운 흐름에 빠르게 대처해야 하는 문화산업에서 비정형 데이터를 분석함으로써 콘텐츠의 활용 및 미래의 트렌드를 예측

2. 대중 가요에 많이 사용되는 키워드를 주제별로 분석

- > 대중문화 형성에 큰 영향력을 가진 K-Pop이 전달하는 메시지를 분석함으로써 대중들이 선호하거나 공감하는 주제 파악

3. 한계점

- > 데이터가 년도 별로 인기가 있었던 20개의 곡으로 이루어져있기 때문에 다양한 토픽의 추출에 실패함(인기가 있었던 곡의 토픽만 뽑아짐) → 더 많은 노래가사를 분석할 필요가 있음