AI DJ

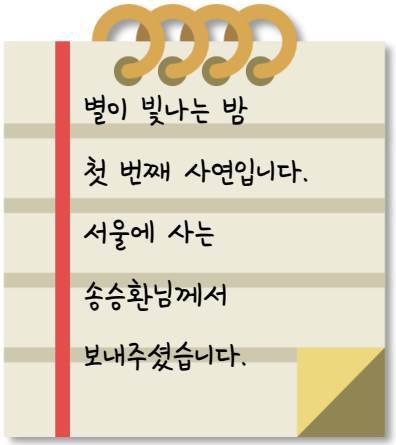
비정형 데이터분석 4조

박건빈 변윤선 이수빈

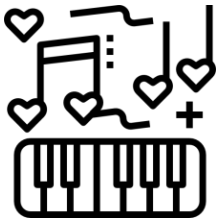
INDEX

01

주제



+  +



→

대숲 사연(Text)기반
음악 추천 시스템 구현

걱정 고민 많은 20대 청춘들이
사연을 많이 올리는 '대나무숲'

진행과정

데이터 수집 ❯ 데이터 정제 ❯ 모델링 ❯ 추천시스템 구현 ❯ 결론

진행과정

데이터 수집 › 데이터 정제 › 모델링 › 추천시스템 구현 › 결론

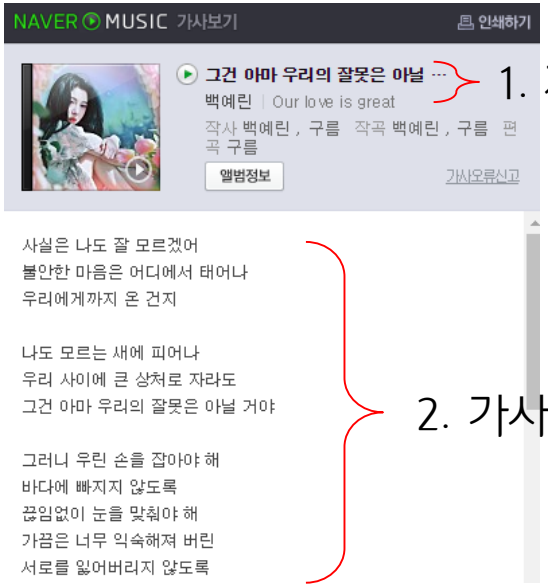
- 음악 정보데이터
- 서울 20개 대학 대나무숲

02

데이터 수집

1) 음악 데이터 수집

-네이버 뮤직 + 벅스 뮤직



1. 제목

2. 가사

3. 테마

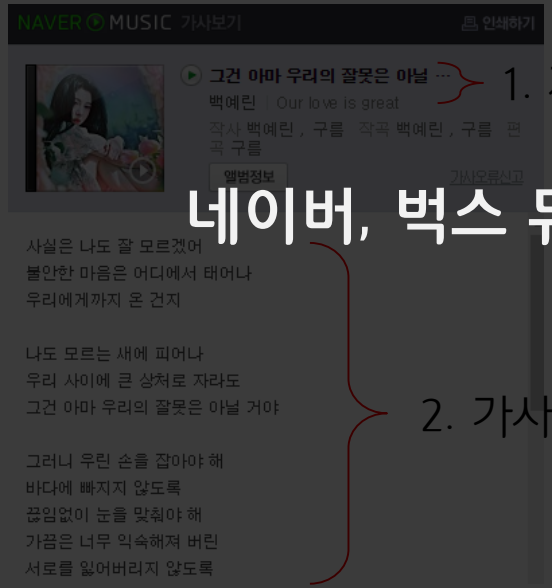
- 국내, 주간 TOP 100, 2016년6월-2019년3월 3째주까지, 약 130주 (130주 X 100=13,000 곡)
- Crawling 항목: 1. 노래제목 2. 노래가사

02

데이터 수집

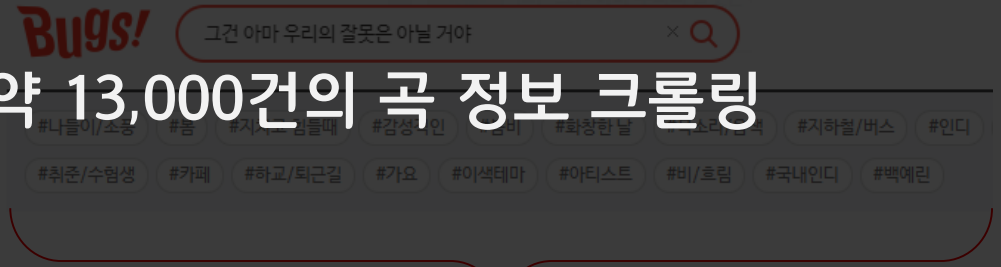
1) 음악 데이터 수집

-네이버 뮤직 + 벅스 뮤직



1. 제목

2. 가사



3. 테마

네이버, 벅스 뮤직에서 약 13,000건의 곡 정보 크롤링

- 국내, 주간 TOP 100, 2016년6월-2019년3월 3째주까지, 약 130주 (130주 X 100=13,000 곡)
- Crawling 항목: 1. 노래제목 2. 노래가사

02

데이터 수집

1) 음악 데이터 수집

곡명 아티스트 가사 — 테마

title	artist	lyric	랩.힙합	휴식.힐링	까페	비오는.날	발라드	R.B.Soul	사랑.설렘	여행.산책
아기공룡	만화천국	아기공룡	0	0	0	0	0	0	0	0
로봇 수사	NA	힘차게 달	0	0	0	0	0	0	0	0
원숭이 (Pe	조권	말썽꾸러기	0	0.2	0	0	0	0	0	0
날아라 슈	NA	치키치키치	0	0	0	0	0	0	0	0
검정고무신	NA	할아버지	0	0	0	0	0	0	0	0
명탐정 코	NA	아침에 눈	0	0	0	0	0	0	0	0
우리의 꿈	NA	내 어린 시	0	0	0	0	0	0	0	0.125
아기공룡	NA	요리보고	0	0	0	0	0	0	0	0
바람의 빛	오연준	사람들만이	0	0.428571	0	0	0	0	0.214286	0.071429
요술공주	NA	너와 나의	0	0	0	0	0	0	0	0
달의 요정	NA	미안해 솔	0	0	0	0	0	0	0	0
꾸러기 수	NA	똘기 뗑이	0	0	0	0	0	0	0	0
천사소녀	NA	오늘밤에	0	0	0	0	0	0	0	0.111111
고요한밤	동방신기	고요한밤	0	0	0	0	0.375	0	0.125	0
글로리아	동방신기	지극히 높	0	0.060976	0.012195	0	0.04878	0	0.02439	0
The First N	동방신기	저 들밖에	0	0	0	0	0.266667	0	0.2	0
크리스마스	러브키즈	크리스마스	0	0	0.2	0	0	0	0.2	0
겨울아이	빅마마	겨울에 태	0	0.047619	0.142857	0	0.238095	0	0.190476	0
소중한 사	길은정	높아만 가	0	0	0	0	0.5	0	0	0

…

드라이브	OST	재즈	겨울연가	봄의왈츠	여름향기	댄스	가을동화	감성	트로트
0	0.333333	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0.272727	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0.272727	0	0	0	0	0	0	0	0
0	0.25	0	0	0	0	0	0	0	0
0	0.222222	0	0	0	0	0	0	0	0
0	0	0	0	0.071429	0	0	0	0	0
0	0.375	0	0	0	0	0	0	0	0
0	0.4	0	0	0	0	0	0	0	0
0	0.25	0	0	0	0	0	0	0	0
0	0.222222	0	0	0	0	0	0	0	0
0	0	0	0.25	0	0	0.125	0	0	0
0.121951	0	0	0.012195	0	0.207317	0.085366	0	0	0
0	0	0	0.266667	0	0	0.133333	0	0	0
0	0	0	0.6	0	0	0	0	0	0
0	0	0	0.285714	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0.166667
0	0.235294	0	0.117647	0	0	0	0	0	0
0	0	0	0.321429	0	0	0.071429	0	0	0
0	0	0	0.2	0	0	0	0	0	0
0	0.083333	0	0	0	0	0.333333	0	0	0

3만곡의 노래 정보

곡명, 아티스트, 가사, 멜론 DJ플레이리스트에 등록 되어있는 테마 데이터를 사용

02

데이터 수집

1) 음악 데이터 수집

곡명 아티스트 가사

테마

title	artist	lyric	랩.힙합	휴식.힐링	까페	비오는.날	발라드	R.B.Soul	사랑.설렘	여행.산책
아기공룡	만화천국	아기공룡	0	0	0	0	0	0	0	0
로봇 수사	NA	힘차게 달	0	0	0	0	0	0	0	0
원숭이 (Pe	조권	말썽꾸러기	0	0.2	0	0	0	0	0	0
날아라 슈	NA	치키치키	0	0	0	0	0	0	0	0
검정고무신	NA	할아버	0	0	0	0	0	0	0	0
명탐정 코	NA	아침에 눈	0	0	0	0	0	0	0	0
우리의 꿈	NA	내 어린 시	0	0	0	0	0	0	0	0.125
아기공룡	NA	요리보고	0	0	0	0	0	0	0	0
바람의 빛	오연준	사람들만	0	0.428571	0	0	0	0	0.214286	0.071429
요술공주	NA	너와 나의	0	0	0	0	0	0	0	0
달의 요정	NA	미안해 솔	0	0	0	0	0	0	0	0
꾸러기 수	NA	똘기 명이	0	0	0	0	0	0	0	0
천사소녀	NA	오늘밤에	0	0	0	0	0	0	0	0.111111
고요한밤	동방신기	고요한밤	0	0	0	0	0.375	0	0.125	0
글로리아	동방신기	지극히 높	0	0.060976	0.012195	0	0.04878	0	0.02439	0
The First N	동방신기	저 들밖에	0	0	0	0	0.266667	0	0.2	0
크리스마스	러브키즈	크리스마스	0	0	0.2	0	0	0	0.2	0
겨울아이	빅마마	겨울에 태	0	0.047619	0.142857	0	0.238095	0	0.190476	0
소중한 사	길은정	높아만 가	0	0	0	0	0.5	0	0	0

드라이브	OST	재즈	겨울연가	봄의왈츠	여름향기	댄스	가을동화	감성	트로트
0	0.333333	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0.272727	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0.222222	0	0	0	0	0	0	0	0
0	0	0	0	0.071429	0	0	0	0	0
0	0.375	0	0	0	0	0	0	0	0
0	0.4	0	0	0	0	0	0	0	0
0	0.25	0	0	0	0	0	0	0	0
0	0.222222	0	0	0	0	0	0	0	0
0	0	0	0.25	0	0	0.125	0	0	0
0.121951	0	0	0.012195	0	0.207317	0.085366	0	0	0
0	0	0	0.266667	0	0	0.133333	0	0	0
0	0	0	0.6	0	0	0	0	0	0
0	0	0	0.285714	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0.166667
0	0.235294	0	0.117647	0	0	0	0	0	0
0	0	0	0.321429	0	0	0.071429	0	0	0
0	0	0	0.2	0	0	0	0	0	0
0	0.083333	0	0	0	0	0.333333	0	0	0

...

과거 멜론에서 크롤링한 약 30,000건의 곡 정보 활용

3만곡의 노래 정보

곡명, 아티스트, 가사, 멜론 DJ플레이리스트에 등록 되어있는 테마 데이터를 사용

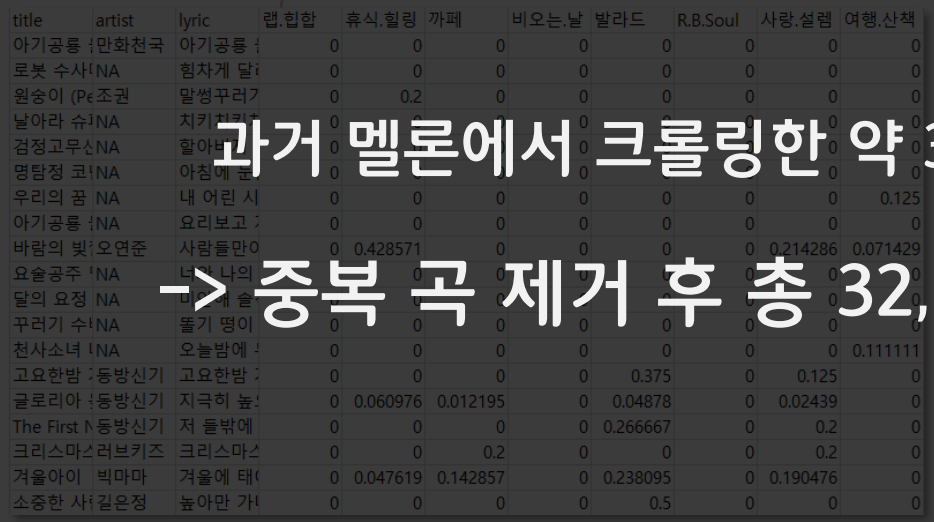
02

데이터 수집

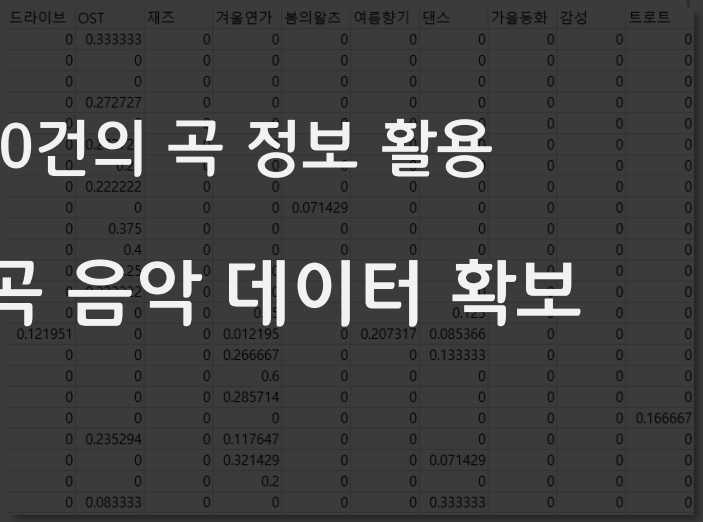
1) 음악 데이터 수집

곡명 아티스트 가사

테마



title	artist	lyric	랩.힙합	휴식.힐링	까페	비오는.날	발라드	R.B.Soul	사랑.설렘	여행.산책
아기공룡	만화천국	아기공룡	0	0	0	0	0	0	0	0
로봇 수사	NA	힘차게 달	0	0	0	0	0	0	0	0
원숭이 (Pe	조권	말썽꾸러기	0	0.2	0	0	0	0	0	0
날아라 슈	NA	치키치키	0	0	0	0	0	0	0	0
검정고무신	NA	할아버	0	0	0	0	0	0	0	0
명탐정 코	NA	아침에 눈	0	0	0	0	0	0	0	0
우리의 꿈	NA	내 어린 시	0	0	0	0	0	0	0	0.125
아기공룡	NA	요리보고	0	0	0	0	0	0	0	0
바람의 빛	오연준	사람들만	0	0.428571	0	0	0	0	0.214286	0.071429
요술공주	NA	너와 나의	0	0	0	0	0	0	0	0
달의 요정	NA	너와 함께	0	0	0	0	0	0	0	0
꾸러기 수	NA	똘기 명이	0	0	0	0	0	0	0	0
천사소녀	NA	오늘밤에	0	0	0	0	0	0	0	0.111111
고요한밤	동방신기	고요한밤	0	0	0	0	0.375	0	0.125	0
글로리아	동방신기	지극히 높	0	0.060976	0.012195	0	0.04878	0	0.02439	0
The First N	동방신기	저 들밖에	0	0	0	0	0.266667	0	0.2	0
크리스마스	러브키즈	크리스마스	0	0	0.2	0	0	0	0.2	0
겨울아이	빅마마	겨울에 태	0	0.047619	0.142857	0	0.238095	0	0.190476	0
소중한 사	길은정	높아만 가	0	0	0	0	0.5	0	0	0



드라이브	OST	재즈	겨울연가	봄의왈츠	여름향기	댄스	가을동화	감성	트로트
0	0.333333	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0.272727	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0.222222	0	0	0	0	0	0	0	0
0	0	0	0	0.071429	0	0	0	0	0
0	0.375	0	0	0	0	0	0	0	0
0	0.4	0	0	0	0	0	0	0	0
0	0.25	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0.129	0	0	0
0.121951	0	0	0.012195	0	0.207317	0.085366	0	0	0
0	0	0	0.266667	0	0	0.133333	0	0	0
0	0	0	0.6	0	0	0	0	0	0
0	0	0	0.285714	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0.166667
0	0.235294	0	0.117647	0	0	0	0	0	0
0	0	0	0.321429	0	0	0.071429	0	0	0
0	0	0	0.2	0	0	0	0	0	0
0	0.083333	0	0	0	0	0.333333	0	0	0

과거 멜론에서 크롤링한 약 30,000건의 곡 정보 활용

-> 중복 곡 제거 후 총 32,000곡 음악 데이터 확보

3만곡의 노래 정보

곡명, 아티스트, 가사, 멜론 DJ플레이리스트에 등록 되어있는 테마 데이터를 사용

02

데이터 수집

2) 사연 데이터 수집

...

4	졸업한지 좀 된 이빨빠진 호랑이에요. 직장생활하면서 쉬는 시간이 생기면 고다
5	여러분 1월 1일은 잘 보내고 계신가요? 고대 학우분께 제 1월 1일을 자랑하고
6	(여성분들 답해주세요!) 안녕하세요. 남자입니다. 제가 어떤여성분을 알게되어서
7	대숲 한달쯤 전에 군대에서 이별했어요. 붙잡고 싶은데 붙잡으면 분명 돌아올 까
8	2017년 1월 1일 0시 0분. 새해가 밝았습니다. 자정에 맞춰둔 알람이 울리고, 우리
9	신년을 맞아 번창하길 바라는 대숲!! 남사친 여사친때문에 논란이 많은데요! 남
10	당신은 항상 나이에 급급해한다. 뒤쳐지고 있는 것 같다고. 하지만 두려워 하지
11	지겨운 방학이 돌아왔다. 기말 전후로 소개팅을 통해 만났던 너 생각에 아무것도
12	여느 때와 없는 날들이다. 평소에 타던 33번 버스는 늘상 그랬듯이 우리 집 앞 정
13	새해에는..... . . . 전역!!!!!!!!! 전역을 하고 싶습니다!!!!!!!!!! 대한민국 군대 다 족구
14	올 호랑이흰님덜,,,,^^ 새해 복 마니 받으시구~^^! 하시는 모든 일들도 번창하시
15	이것은 유언장이 아니다 며칠에 걸친 지리한 싸움이 일단락됐다 어디서부터 잘
16	11시 45분. '이제 슬슬 올라가봐야하지않나?' 곧 도시 곳곳에서 피어날 불꽃들을
17	축하한다. 여러분은 벌써 올해의 1/365을 성공적으로 낭비했다. 대학생 (0/1/0)
18	즐거운 마음으로 참석한 새터(혹은 개강파티)... '나는 1X학번인데, 편하게 대해도
19	몇년째 엄청 친한 여사친인데 한때 짝사랑의 대상이었지만 마음을 깔끔히 접었
20	어째서 교수님께서는 아직까지 성적을 공개하지 않으시는 걸까. 우리를 위한 나
21	고대 대나무숲 지기들 닉넴을 ~냥으로 통일했던데..... 고양이인거 ㅇㅈ하는 부분
22	노트북으로 드라마 보는데 화면 어두워질 때 비치는 내 모습이 존못이라 볼 때마
23	가끔 절 이유 없이 싫어하는 사람들이 있어요. 그러면 저는 그 사람들에게 이 악
24	정유년은 붉은 닭의 해입니다. 그렇습니다. 양념치킨의 해라는 거죠.
25	님들 솔직히 비쁠보다 그냥 에이가 더 기분 나쁘지않아요? 아 이번 학기 망했어
26	어둠은 빛을 이길 수 없다... 2017년 새해가 밝았습니다. 어둠은 조그마한 빛만
27	글을 잘 쓰고 싶다. 이과생인 내가 이런 재능을 탐하는 건 도둑놈 심보라는 걸 알
28	나는, 간질입니다 여러분이 생각하는 그 간질은 아니에요 막 거품물고 쓰러져서
29	제가 군대 가는 것 때문에 여자친구가 너무너무너무너무 많은 걱정을 해서 고민

서울 시내 20개 대학에서 각 2000개 씩 크롤링

03

데이터 수집

2) 사연 데이터 수집

가천대학교 대나무숲

@gcubamboo

- 홈
- 정보
- 사진
- 동영상
- 게시물
- 커뮤니티

한양대학교 대나무숲

@hyubamboo

- 홈
- 정보
- 사진
- 동영상
- 게시물
- 커뮤니티

… 사연 데이터 약 40000만개 확보 →

4	졸업한지 좀 된 이빨빠진 호랑이에요. 직장생활하면서 쉬는 시간이 생기면 고디
5	여러분 1월 1일은 잘 보내고 계신가요? 고대 학우분께 제 1월 1일을 자랑하고
6	(여성분들 답해주세요!) 안녕하세요. 남자입니다. 제가 어떤여성분을 알게되어서
7	대숲 한달쯤 전에 군대에서 이별했어요. 붙잡고 싶은데 붙잡으면 분명 돌아올 까
8	2017년 1월 1일 0시 0분. 새해가 밝았습니다. 자정에 맞춰둔 알람이 울리고, 우리
9	신년을 맞아 번창하길 바라는 대숲!! 남사친 여사친때문에 논란이 많은데요! 남
10	당신은 항상 나이에 급급해한다. 뒤쳐지고 있는 것 같다고. 하지만 두려워 하지
11	지겨운 방학이 돌아왔다. 기말 전후로 소개팅을 통해 만났던 너 생각에 아무것도
12	여느 때와 없는 날들이다. 평소에 타던 33번 버스는 늘상 그랬듯이 우리 집 앞 정
13	새해에는..... . . . 전역!!!!!!!! 전역을 하고 싶습니다!!!!!!!!!! 대한민국 군대 다 족구
14	울 호랑이흰님덜,,,,^^ 새해 복 마니 받으시구~^^! 하시는 모든 일들도 번창하시
15	이것은 우연자이 아니다 면치에 걸친 지리한 싸움이 일단락됐다 어디서부터 잘
16	않나?' 곧 도시 곳곳에서 피어날 불꽃들을
17	성공적으로 낭비했다. 대학생 (0/1/0)
18	즐거운 마음으로 참석한 새터(혹은 개강파티)... '나는 1X학번인데, 편하게 대해도
19	몇년째 엄청 친한 여사친인데 한때 짝사랑의 대상이었지만 마음을 깔끔히 접었
20	어째서 교수님께서는 아직까지 성적을 공개하지 않으시는 걸까. 우리를 위한 나
21	고대 대나무숲 지기들 닉넴을 ~냥으로 통일했던데..... 고양이인거 ㅇㅈ하는 부분
22	노트북으로 드라마 보는데 화면 어두워질 때 비치는 내 모습이 존못이라 볼 때마
23	가끔 절 이유 없이 싫어하는 사람들이 있어요. 그러면 저는 그 사람들에게 이 악
24	정유년은 붉은 닭의 해입니다. 그렇습니다. 양념치킨의 해라는 거죠.
25	님들 솔직히 비쁠보다 그냥 에이가 더 기분 나쁘지않아요? 아 이번 학기 망햇어
26	어둠은 빛을 이길 수 없다... 2017년 새해가 밝았습니다. 어둠은 조그마한 빛만
27	글을 잘 쓰고 싶다. 이과생인 내가 이런 재능을 탐하는 건 도둥놈 심보라는 걸 알
28	나는, 간질입니다 여러분이 생각하는 그 간질은 아니에요 막 거품물고 쓰러져서
29	제가 군대 가는 것 때문에 여자친구가 너무너무너무너무 많은 걱정을 해서 고민

서울 시내 20개 대학에서 각 2000개 씩 크롤링

진행과정

데이터 수집 > 데이터 정제 > 모델링 > 추천시스템 구현 > 결론

-맞춤법 검사
-불용어 처리, tokenize

03

데이터 정제

1) 맞춤법 검사

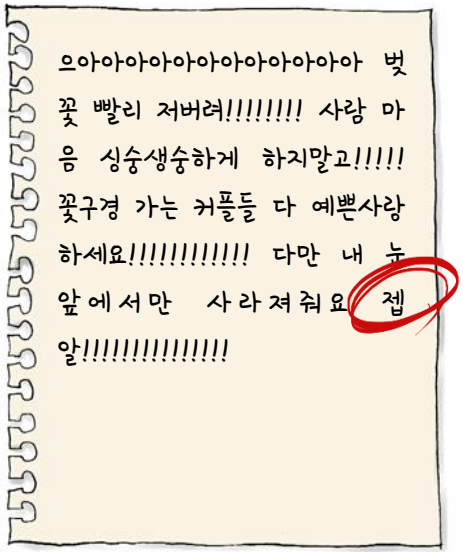
으아아아아아아아아아아아아 벚꽃 빨리 저버려!!!!!!!! 사랑 마음 싱숭생숭하게 하지말고!!!!! 꽃구경 가는 커플들 다 예쁜사랑 하세요!!!!!!!!!!!! 다만 내 눈앞에서만 사라져줘요 젭알!!!!!!!!!!!!!!!

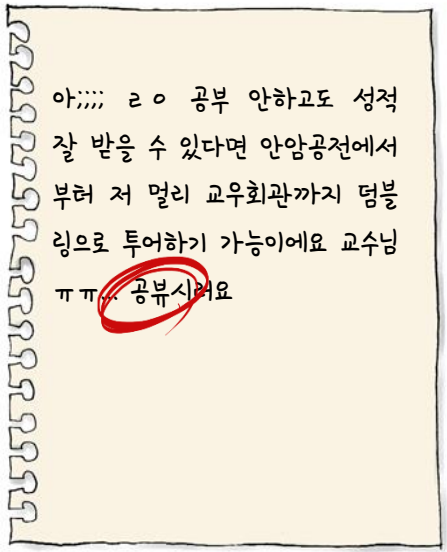
아;;;; ㄹㅇ 공부 안하고도 성적 잘 받을 수 있다면 안암공전에서부터 저 멀리 교우회관까지 덤블링으로 투어하기 가능이에요 교수님 ㅠㅠ... 공뷰시러요

03

데이터 정제

1) 맞춤법 검사

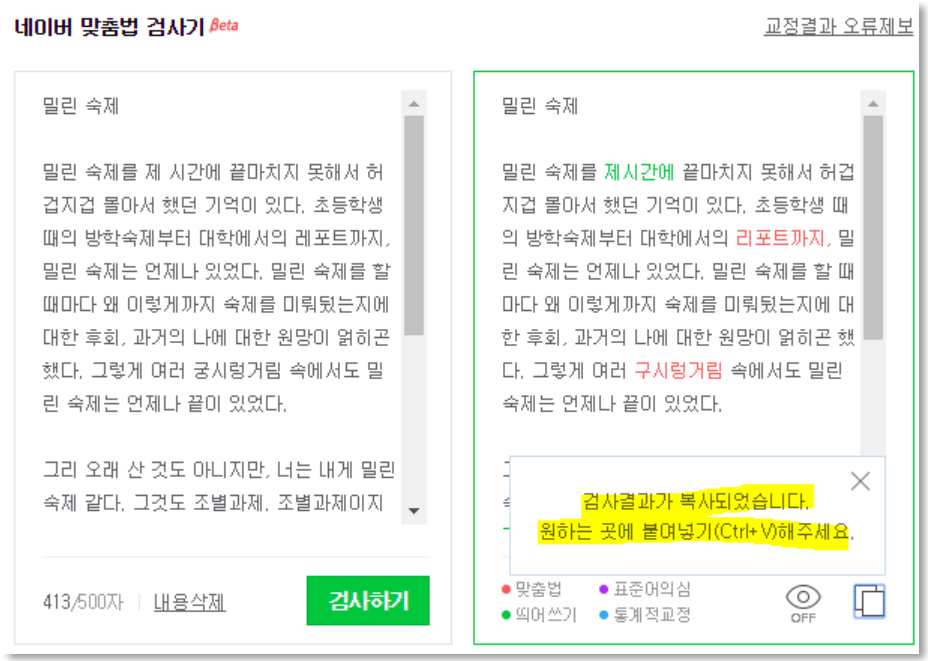




03

데이터 정제

1) 맞춤법 검사



네이버 맞춤법 검사기로 맞춤법 확인
Python selenium 이용

03

데이터 정제) Tokenize

#2430번_제보 어제 관정 열람실에서 11시까지 공부하다가 깜빡하고 좌석 반납을 안했더니 바로 오늘부터 열흘 간 이용 정지 먹었어요.. 도서 반납도 30일까지는 연체가 가능한데 한 번 반납 안하고 집 갔다고 바로 이용 정지 맥이나요ㅠ 당장 중간고사가 코앞인데 너무 빡이 치네요.. 도서관 이용하실 때 깜빡하지 맙시다„

유니코드, 문장부호, 숫자, 영어, 한자, 그 외 불용어 제거

"['어제', '관정', '열람', '실', '시', '공부', '깜빡', '좌석', '반납', '안', '오늘', '열흘', '간', '이용', '정지', '먹다', '도서', '반납', '까지는', '연체', '가능하다', '반납', '안', '집', '가다', '이용', '정지', '맥', '나오다', '중간고사', '코앞', '너무', '빡', '치네다', '도서관', '이용', '깜빡', '맙시']"

Tokenize 완료

2) Tokenize

#2430번_제보 어제 관정 열람실에서 11시까지 공부하다가 깜빡하고 좌석 반납을 안했더니 바로 오늘부터 열흘 간 이용 정지 먹었어요.. 도서 반납도 30일까지는 연체가 가능한데 한 번 반납 안하고 집 갔다고 바로 이용 정지 맥이나요ㅠ 당장 증간고사가 코앞인데 너무 빡이 치네요.. 도서관 이용하실 때 깜빡하지 맙시다..

유니코드, 문장부호, 숫자, 영어, 한자, 그 외 불용어 제거

총 1,600개

불용어 사전 오픈 데이터(1400개) + 직접 추가(200개)

"['어제', '관정', '열람', '실', '시', '깜빡', '좌석', '반납', '안', '오늘', '열흘', '간', '이용', '정지', '먹다', '도서', '반납', '까지는', '연체', '가능하다', '반납', '안', '집', '가다', '이용', '정지', '맥', '나오다', '중간고사', '코앞', '너무', '빡', '치네다', '도서관', '이용', '깜빡', '맙시']"

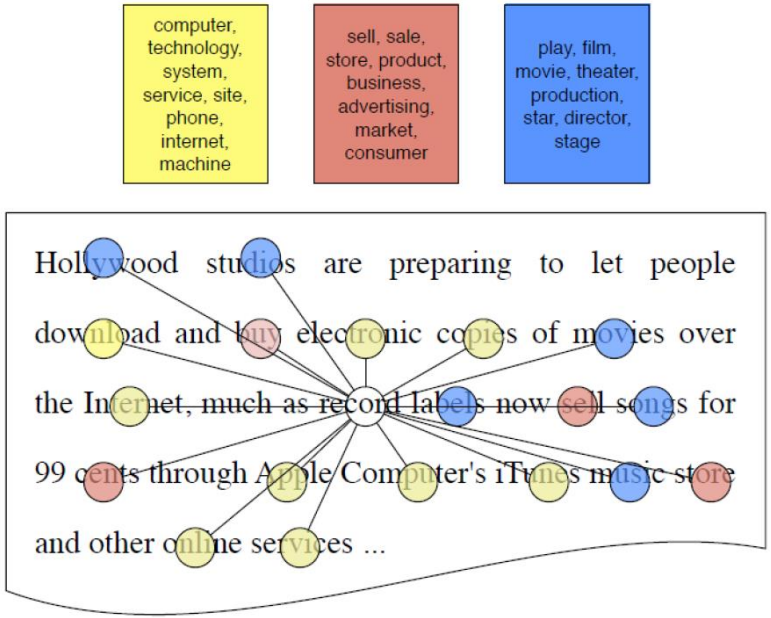
Tokenize 완료

진행과정

데이터 수집 › 데이터 정제 › 모델링 › 추천시스템 구현 › 결론

-토픽 모델링
-임베딩

1) LDA



사연 데이터

['어제', '관정', '열람', '실', '시', '공부', '깜빡', '좌석', '반납', '안', '오늘', '열흘', '간', '이용', '정지', '먹다', '도서', '반납', '까지는', '연체', '가능하다', '반납', '안', '집', '가다', '이용', '정지', '맥', '나오다', '중간고사', '코앞', '너무', '빡', '치네다', '도서관', '이용', '깜빡', '맙시']

음악 테마 데이터

100시리즈 2010년대 베스트 가요 감성적인 쓸쓸한 외로울때 발라드 발라드한 2018년 가을 송년회 늦가을 쌀쌀한 날 이별/슬픔 노래방 가요 가을밤 연도별음악 연도별 인기가요 2010년대 뮤직기네스

04

모델링

1) LDA - 사연

LDA에서 얻은 키워드 기반,
9개의 topic으로 분류

Manually Naming

- 2+5: 일상/우울
- 4: 일상/밝은/학교
- 6: 새벽 감성
- 7+18: 일상/학교
- 8: 사랑/슬픔
- 9+11: 가족
- 10:사회/답답함
- 16+19: 학교/우울
- 17: 사랑/설렘

Topic 수 = 20

개	적	먹다	돈	분	군대	아프다	학교	내	글	여성	아버지	오빠	노래	사람	애인	수업	안	가다	말다
두	인	밥	원	시	군	오늘	분들	사랑	엄마	수강신청	인간	사진	부르다	없다	예바	교수	너무	공부	술
자리	되다	귀	소리	안녕하다	한국	눈	학생	좋아하다	아빠	정부	충	여행	기숙사	않다	문	듣다	친구	돌다	선배
이름	한	아침	게임	새내기	전역	꿈	대	마음	읽다	집단	운동	찍다	곡	생각	옐시	담배	좋다	집	제발
않다	대해	인가요	방	궁금하다	입대	세상	동아리	난	쓸다	우리나라	어머니	영화	세계	보다	열다	냄새	싫다	나오다	욕
세	문제	물	팀	계시다	번만	밤	기간	만나다	대숲	범죄	이든	선물	환경	말	마르다	짜다	아니다	시험	당하다
성대	아니다	뜨다	할머니	친해지다	군인	속	되다	다시	찾다	투표	주의	누르다	변화	못	내려가다	학부	그냥	오다	종강
휴학	대한	마시다	비	학번	하고도	울다	학우	날	댓글	검색	아들	카페	영상	되다	성매매	피우다	어떻다	열심히	인사
지키다	생각	저녁	넣다	후배	뉴스	손	많다	나르다	올리다	인권	사고	장	맺다	그렇다	침대	조	보다	다니다	씨
곳	않다	돌리다	대체	새	살리다	길	생	이제	쓰다	회의	여름	주고받다	영어	싶다	글자	강의	전	부모님	술자리
중	위	맛있다	풀	가요	복무	잠	드리다	헤어지다	써다	뽑다	죽음	꽃	보상	더	섹스	어색하다	진짜	대학	본인
지원	경우	음식	천	전공	팬	옷다	질문	행복하다	입다	북한	폭력	크리스마스	시끄럽다	모르다	역	화장실	랑	학기	불편하다
번째	되어다	병	모으다	들다	인턴	하루	단	미안하다	올라오다	자유	무조건	잘해주다	볼다	아니다	공연	파이팅	돌다	중	인간관계
실	성	기울이다	비싸다	티	기업	눈물	점	나다	옷	대표	키우다	떨다	만족	좋다	완벽하다	토	자다	받다	빌리다
확인	위해	맛	히	공	재	모습	서울	거야	해봤다	정치	긍정	끊기다	전문	한	순위	왜냐면	그렇다	정도	장난
도서관	받다	참치	발권	분위기	리	기억	활동	지내다	달다	대한민국	시선	빨다	가사	지금	세다	술집	돼다	준비	무시
이던	사회	쳐다보다	바닥	입학	청춘	오다	감사하다	넌	동생	나라	드립	식다	출석	많이	가도	옷음소리	근데	생활	톡
사용	가지	심장	올리다	터	두산	순간	타	서로	서강대	지지	죽이다	불법	탑	힘들다	유독	건물	얘기	나가다	죄송하다
옆	때문	소주	비용	흑	절반	목소리	관	넓	남기다	페미니즘	연예인	프로필	만족하다	정말	큰일	이제야	싫다	돌어가다	들이다
음악	어떻다	사주다	다리	배	나라	이별	가능하다	잊다	최고	국가	닳다	그러다가	유튜브	걸	프로	현명하다	이렇다	돌어오다	조용하다
성희롱	이유	돌다	백	과목	의경	떠오르다	이나	감정	딸	통합	죄	드디어	태양	많다	무게	강의실	연락	놀다	줄다
버스	의견	핸드폰	과외	쯤	외국	가슴	장학금	아직도	억지로	교내	역사	아른거리다	이자	자다	스러운	청소	대숲	학년	끼리
렵다	이렇다	켜다	롬메	일로	작성	그녀	도움	지나다	감사하다	공지	가족	비밀	특징	해주다	범죄자	볼	여자친구	누나	면
나무숲	개인	잔	지르다	다녀오다	근무	지나가다	중	연애	꾸준하다	종교	썩다	매달리다	여지	그렇게	자유롭다	너무하다	남자친구	끝나다	거리
임	내용	끼	소문	성균관대	뮴다	예쁘다	겁니다	슬프다	색	대표자	에야	싸	갑	알	살인	캠퍼스	남자	후	거뮬다
바꾸다	부분	힘내다	고양이	길이	그다지	울	하나요	기억	겁	근거	장이	섭섭하다	노래방	이렇게	성관계	디도	오다	갈다	식
공간	이라는	아아	신고	컴퓨터	베치	그리다	관련	처음	달리다	주의자	으른	필름	관점	주다	전날	시점	건	고등학교	개강
선수	이해	시도	아저씨	신청	부대	어제	학생회	보고	페이스북	부스	엘리베이터	막말	훈련	받다	이불	김	보고	주	상관없다
부탁드리	입장	가게	연락처	내년	병사	내일	관심	연락	구체	강제	동동하다	매물	지식	때문	러	차례	알다	곳	진지하다
돌아다니	해결	영	양	화석	장점	별	관리	끝	자랑스럽다	일본	결코	택시	잃어버리다	알다	안동	딜	결다	살	뒤
소	책	급하다	생리	여주다	직업	작다	필요하다	마지막	종류	올림픽	석	쥐	시가	사실	자제	감독	여자	반	달라

04

모델링

1) LDA - 음악 테마

LDA에서 얻은 키워드 기반,
9개의 topic으로 분류

Manually Naming

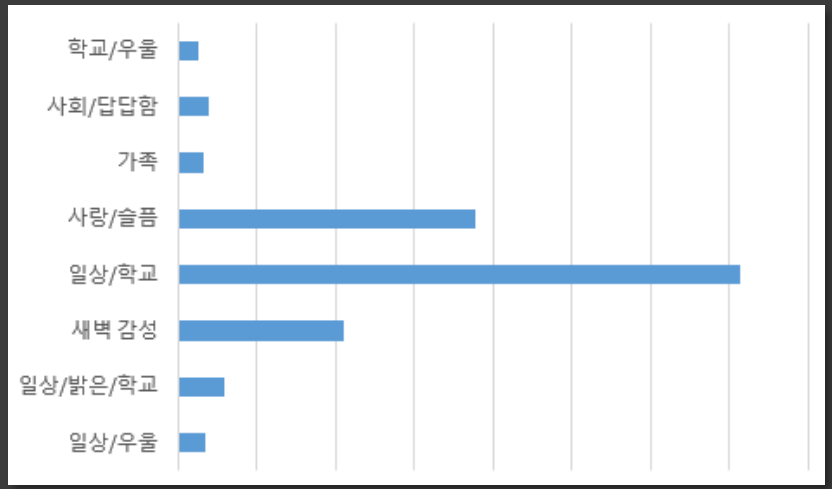
1: 일상/밝은
2: 사랑/설렘
3+17: 힙합
7: 힐링
8+12: 기분전환
11: 새벽 감성
14: 사랑/슬픔
15: 일상/위로
19: 가족

Topic 수 = 20

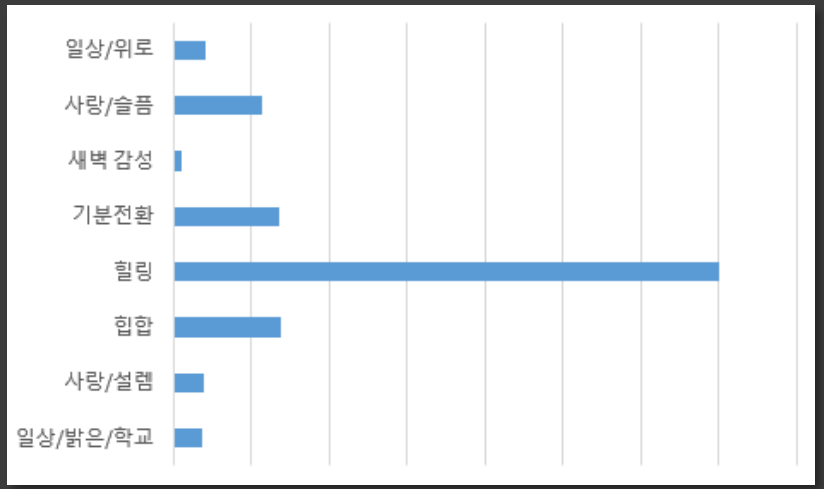
그녀	크다	뽐뽐뽐	연애	뉴잭스윙	꿈	봄	운명	안녕	용기	눈물	속	아침	비	무례하다	내	날	따르다	노래	온
여자	거기	잭슨	고독	별놈	세상	행복	영화	나이	야생마	잊다	거짓말	벚꽃	내리다	만세	사랑	더	추다	라라라	먹다
놀다	반	진진	봄비	생마	하늘	일상	인연	커피	탱탱	아프다	진실	생	바다	뻑	보다	전	줌	나나	집
엄마	달리다	빙그르	피자	다가닥다가	위	키스	버스	노을	단둘	기억	억	눈뜨다	여름	엽	없다	위로	음악	라라	꽃
가져가다	그림자	계리	빠라빠빠	뽑	빛	주인공	저녁	연인	서울	떠나다	드라마	토	파도	갈매기	난	겨울	몸	말라	돈
오오	맛	계릭	초록	마야	소리	색깔	힙합	정	낡다	추억	안개	이마	아아	죽하	맡	인생	크리스마스	랩	피다
오빠	뚝	꿀잼	매타	얄리	만의	맥주	따따	오후	방식	흐르다	추악하다	마리	비밀	부릉	그대	고백	타	랄라라	학
개	그걸	마이클	노노	빠빠	꾸다	봄바람	헬로	창가	샤	그립다	거짓	우와	시원하다	지구인	않다	그리움	즐기다	라랄	지리
뺏다	볼	흐드드	미랠	앞다리	작다	데이트	세종대왕	분만	두두	이별	워워워	부	뚜	빙고	널	미련	히	멜로디	잔
아빠	올리다	이어폰	나즈	팃다리	뜨다	생애	파티	카페	산소	슬프다	닫다	밉다	보아	신속하다	못	이야기	빠르다	기울이다	출쳐
리	원	세뇨리따	코코	감성	뜨겁다	무비	장면	찰칵	네모	지우다	또다른	버려진	우산	정기석	니	가을	리듬	빠빠빠	배
노	머	강자볕	노노노	뚜뚜뚜루	달다	팬	길가	케익	고향	울	가리우다	톡톡	루루	옅다	맘	가요	희망	랄랄	우주
사로자다	세우다	더랬	얄밉다	뚭뚜	새	해피	카카오	굿다	두르다	보내다	술	우유	빗	덤벼	싶다	편지	야야	초콜릿	자장
그녘	질문	이정표	햄	뚭뚜루	태양	커플	회색	결혼	헬헬	울다	높	자전거	예감	도장	나르다	청춘	주문	대중	빨
반하다	군	염소	케나	홀러러러러	구름	한강	돛대	집중	애매하다	떠나가다	파란하늘	체리	바다로	잼	내다	번만	키	뚜뚜루	맛있다
빠다	관심	아라	쿠라	꿈는다	눈부시다	사랄	맞팔	지랄	쿠바	잊혀지다	서랍	리오	두비	된장국	모르다	전화기	소	슈가	다리
찌릿	분	마카레나	밤밤	레베카	아래	마카롱	정류장	모금	급	아픔	워깊	하아	뚜뚜뚜	스크림	그럴다	첫사랑	코	마마	술잔
근데	감	내디디다	숙제	위고	달려가다	잼잼	블루스	로맨틱	이틀	슬픔	울결	재다	덥다	땡	사람	낙엽	빨갛다	도비	검다
생일	톡	복일	코노	으랴거친	날개	익	짜이다	맴	나머지	상처	외침	죽음	무언가	제자	오다	활	춤추다	흥얼거리	빨다
르	일으키다	특급	레조	나도야	타고	둘썩들썩	화이트	기차	자신감	남아	파도처럼	맨날	바닷가	겨려	한	온대	동네	내달	배달
꺼	에게도	흐드	삐쭉	오라버니	밝다	흰색	숭	공원	운	흘리다	흐름	나니	여름밤	삼포	되다	집착	맡기다	콧노래	스르르
우후	산	드타	이키	스파이더맨	꿈꾸다	베이스	발표	샌드위치	아리랑	돌아서다	오래되다	모노	그치다	세기	눈	가기	도록	재미	소재
홈런	어차피	문워크	샤라	쉬엄쉬엄	크게	겁니다	우연이	슈퍼	과녁	미우다	묶이다	음식	섬	할머니	걸	언	올해	세레나데	택시
홈	커	엽싸다	아루	까페	날다	포옹	자유로	부모님	깃발	세월	혼돈	미미	오우	수위	마음	힐	잘생기다	마음껏	금도
남자친구	형	순정	우시	만쥬	높다	호빵	셀프	졸리다	라인	지워지다	나비야	다미	쿨하다	살라	좋다	고단	짜릿하다	무지	게임
아가씨	제대로	트윈스	치카	나이팅게일	지르다	그랬는데	존	지영	비극	뒷모습	이니셜	지우	비처럼	허풍	넌	어쨋밥	윤직이다	슈비	스르륵
왼전하다	우울하다	팔로우	자욱	중화반점	일어나다	도현님	앨볌런스	래퍼	매혹	와의	무뎌져	아우	푸른	산이	다시	소나기	비밀스럽	년대	취하
오케이	담	일대	연주	미사일	높이	쿵짝지	평화	깨깍	강	다짐	흑백	꼼찍하다	뚜뚜	출근길	버리다	가족	모이다	무어	강
오리	학교	락앤롤	아시	가시리	오오오	잠결	영화관	와인	주르르	시리다	슾숙	메롱	여인	외계인	가다	나인	나노	발차기	짝
잡초	적당하다	구멍이	레노	스트레스	문	이모티콘	스냅백	일요일	대한민국	어도	락자	요리	루	엘리베이터	너무	하늘아	홀쩍	출구	부풀다
소개	주말	우어	파파	홀러러러러	별빛	담벼락	스나이퍼	불쑥	쿵짝	굿바이	성당	즈	루비	장난감	안	점음	분위기	기막히다	달달

1) LDA - 토픽 분포

사연 Topic



음악 테마 Topic



2) Doc2Vec

사연 2 + 5	일상/우울	노래 15	일상/위로
사연 4	일상/밝은/학교	노래 1	일상/밝은
사연 6	새벽 감성	노래 11	새벽 감성
사연 7 + 18	일상/학교	노래 8 + 12	기분전환
사연 8	사랑/슬픔	노래 14	사랑/슬픔
사연 9 + 11	가족	노래 19	가족
사연 10	사회/답답함	노래 3 + 17	힙합
사연 16 + 19	학교/우울	노래 7	힐링
사연 17	사랑/설렘	노래 2	사랑/설렘

Pair별 embedding 공간 생성 (Doc2Vec)

2) Doc2Vec

Original Data

왜냐면 난 걔를 좋아해 걔가 어떤 앤지 알고서도 좋아해 걔가 다른 남잘 만나는 걸 알고 있지 나는 너는 여신이니까
그래도 돼 너는 여신이니까 여신이니까 여신이니까 너는 되고 나는 다 안 되는 게 법이니까 난 그냥 널 봐서
강아지처럼 기분이 좋아 귀여워할 때만 나를 사용해도 좋아 호구 기리보이

Tokenized Data

"['왜냐면', '난', '걔', '좋아하다', '걔', '어떻다', '앤지', '알다', '좋아하다', '걔', '남다', '만나다', '걸', '알', '여신', '이니까', '돼다', '여신', '이니까', '여신', '이니까', '여신', '이니까', '되다', '안', '되다', '법', '이니까', '난', '그냥', '널', '보다', '강아지', '기분', '좋다', '귀엽다', '나르다', '사용', '좋다', '호구', '기리보이']"

```
array([ 0.6298259 , -0.09085082, -0.28519925,  0.5714372 ,  0.25872275,
        0.60149074, -0.02625632,  0.07765006, -0.7839311 , -1.5680498 ,
       -0.92545795,  0.13430485,  0.36699116,  0.15147817, -0.96257496,
       -0.78277445,  0.5164371 ,  0.29842624, -0.205785  ,  0.22466424,
       -0.9545996 ,  0.6128009 , -0.349194  , -0.31204844, -0.7457065 ,
       -0.54703647,  0.1896811 ,  0.1491383 , -0.21775162, -0.86559105,
       -0.48137563, -0.16984546,  0.8141323 ,  0.0564562 ,  1.0037832 ,
        0.81033397, -0.58311844,  0.5160924 ,  0.510567  , -0.5937158 ,
        0.5981595 ,  0.87641585, -0.5774979 , -0.40876126,  0.3795267 ,
       -0.4001868 ,  0.07840311, -0.7881859 ,  0.22949971,  0.42853683],
      dtype=float32)
```

2) Doc2Vec

Original Data

왜냐면 난 걔를 좋아해 걔가 어떤 앤지 알고서도 좋아해 걔가 다른 남잘 만나는 걸 알고 있지 나는 너는 여신이니까
그래도 돼 너는 여신이니까 여신이니까 여신이니까 너는 되고 나는 다 안 되는 게 법이니까 난 그냥 널 봐서
강아지처럼 기분이 좋아 귀여워할 때만 나[illegible]보이

총 80,000개의
사연, 가사 데이터에 대해
9개의 space에 embedding

Tokenized Data

"['왜냐면', '난', '걔', '좋아하다', [illegible], '앤지', [illegible], '좋아하다', '걔', '다른', '남잘', '만나다', '걸', '알', '여신', '이니까', '돼다', '여신', '이니까', '여신', '이니까', '여신', '이니까', '되다', '난', '되다', '법', '이니까', '난', '그냥', '널', '보다', '강아지', '기분', '좋다', '귀엽다', '나르다', '사용', '좋다', '호구', '기리보이']"

```
array([ 0.6298259 , -0.09085082, -0.28519925,  0.5714372 ,  0.25872275,
        0.60149074, -0.02625632,  0.07765006, -0.7839311 , -1.5680498 ,
       -0.92545795,  0.13430485,  0.36699116,  0.15147817, -0.96257496,
       -0.78277445,  0.5164371 ,  0.29842624, -0.205785  ,  0.22466424,
       -0.9545996 ,  0.6128009 , -0.349194  , -0.31204844, -0.7457065 ,
       -0.54703647,  0.1896811 ,  0.1491383 , -0.21775162, -0.86559105,
       -0.48137563, -0.16984546,  0.8141323 ,  0.0564562 ,  1.0037832 ,
        0.81033397, -0.58311844,  0.5160924 ,  0.510567  , -0.5937158 ,
        0.5981595 ,  0.87641585, -0.5774979 , -0.40876126,  0.3795267 ,
       -0.4001868 ,  0.07840311, -0.7881859 ,  0.22949971,  0.42853683],
      dtype=float32)
```

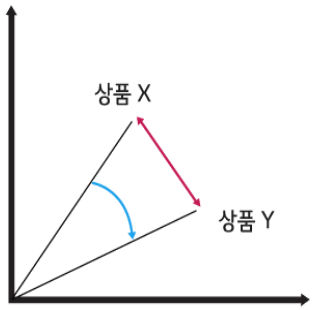
진행과정

데이터 수집 › 데이터 정제 › 모델링 › 추천시스템 구현 › 결론

-코사인 유사도 기반의
음악 5곡 추천

모델링

코사인 유사도 기반



$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

사연과 유사도 높은 Top 5 음악

	artist	title	lyrics
1	로시(Rothy)	다 핀 꽃	['핀', '이름', '모르다', '꽃', '송이', '떠나다', '전의', '끝'...
	산이(San E)	나 왜이래 (feat. 강민희 Of 미스에스)	['널', '바라보다', '표정', '관리', '안', '돼다', '얼굴', '써다...
⋮	임정희	Golden Lady (feat. 현아 Of 4Minute)	['없이', '어떻다', '살아가다', '바보', '질문', '말', '알', '남...
	먼데이 키즈 (Monday Kiz)	가을 안부	['어떻다', '자다', '지내다', '지난', '여름', '유난히', '힘', '...
5	에디킴(Eddy Kim)	긴 밤이 오면	['기다', '밤', '오다', '길', '잃다', '니', '없다', '난하다',...

진행과정

데이터 수집 › 데이터 정제 › 모델링 › 추천시스템 구현 › 결론

05

결과

교수님이 어린이날에 보강을 한다네요. 휴일이라 친구들이랑 여행 계획도 짜고 예약까지 해놨는데 (···) 굳이 휴일에 수업을 해야하나요.

Play List

Title	Artist
Dangerous	보아 (BoA)
Pass	쥬얼리
A-G-E	E SENS
Tokyo Inn	혁오
Lonely(없구나)	B1A4

05 결과

교수님이 어린이날에 보강을 한다네요. 휴일이라 친구들이랑 여행 계획도 짜고 예약까지 해놨는데 (···) 굳이 휴일에 수업을 해야하나요.

추천 곡 1
(Dangerous- 보아)

··· 장난치지 말고
속이려 들지 말고
서투른 변명 말고
Hurry up hurry up ···

··· 앞뒤 맞질 않지
넌 내손안에 있지
넌 참 뻔하게 보이는
습관적 거짓말 늘어놔···

추천 곡 2
(Pass- 쥬얼리)

··· 이젠 PASS 너는 PASS
참을만큼 참았단 말야
답답해 지루해
늘 네 기준에 날 맞추고···

05

결과

조현아 동영상을 보려고 했다.
끝까지 보지 못하고 중간에 꺼야
했다. 엄마가 소리 지르는 모습이
오버랩 되었기 때문이다. (···)
그렇다고 엄마가 날 사랑하지 않은
건 아니다. (···)엄마가 없었으면
지금의 나는 없었을 거다.

Play List

Title	Artist
엄마로 산다는 것은	보아 (BoA)
세노야	쥬얼리
별 헤는 밤	E SENS
그 아버지에 그 아들	혁오

05

결과

선배 저한테 왜 그러셨어요.(···) 저
애정결핍 얼마나 심각한지
모르시잖아요. 문득 떠오르는 조각들이
저를 얼마나 힘들게 하는지
아시나요(···)그냥 마음 없던 척
선배 모르던 척 살아갈래요.

Play List

Title	Artist
헤어지기로 해	넬 (NELL)
Idiot	카라
연애 같은 걸 하니까	소란 (SORAN)

05 비교

1) 네이버 뮤직

교수님이 어린이날에 보강을 한다네요. 휴일이라 친구들이랑 여행 계획도 짜고 예약까지 해놨는데 (···) 굳이 휴일에 수업을 해야하나요.

검색 키워드

보강

Play List

TITLE	ARTIST
고급스런 명상 뉴에이지 (건강보강, 신경피로 완화)	일상의 휴식
늦은 후회	Viloet F

2) 라디오 사연

대학 신입생입니다. 저는 시골에 살아 소규모 학교를 다녔습니다. 그래서 친구들끼리 우정도 깊고 친밀했습니다. (···) 친구 그영이 원하는 대학에 합격하지 못해 아직 공부 중입니다. 다시 모여서 재밌게 놀자! 파이팅 !

로이킴, 정준영의 친한친구 사연 中

신청곡

슈퍼스타 – 이한철

Play List

TITLE	ARTIST
혜화동 (혹은 쌍문동)	박보람
zero	백예린

Summary

추천 LOGIC

1. 대나무숲 사연 데이터 입력
2. 데이터 정제: 맞춤법 검사, 불용어(stopword) 제거, Tokenization
3. LDA로 나눈 9개의 topic 중 가장 가까운 topic으로 분류
4. 해당 topic과 matching 시켜 놓은 노래 카테고리의 임베딩 공간에서 Input data 임베딩
5. 코사인 유사도가 높은 노래 5개를 output으로 추출

개선 사항

멜로디 정보를 Feature로 추가하면, 더 좋은 결과를 얻을 것으로 기대

Reference

Stopwords open data

- https://github.com/stopwords-iso/stopwords-ko
- https://www.ranks.nl/stopwords/korean
- http://blog.daum.net/_blog/BlogTypeView.do?blogid=0FyGV&articleno=7611081&categoryId=223515&regdt=20080526063911

감사합니다