

댓글 분석을 통한



나만의 플레이리스트

박 선 경

목 차

01

개발 환경

02

주제

03

데이터 수집

04

데이터 분석

05

HTML

06

웹시연

























개 발 환 경

항목	상세내용
O/S	✓ window
언어	✓ PYTHON, HTML, CSS
라이브러리	✓ pandas, numpy, selenium,bs4,MariaDB,konlpy,flask

주제

총 1,324개

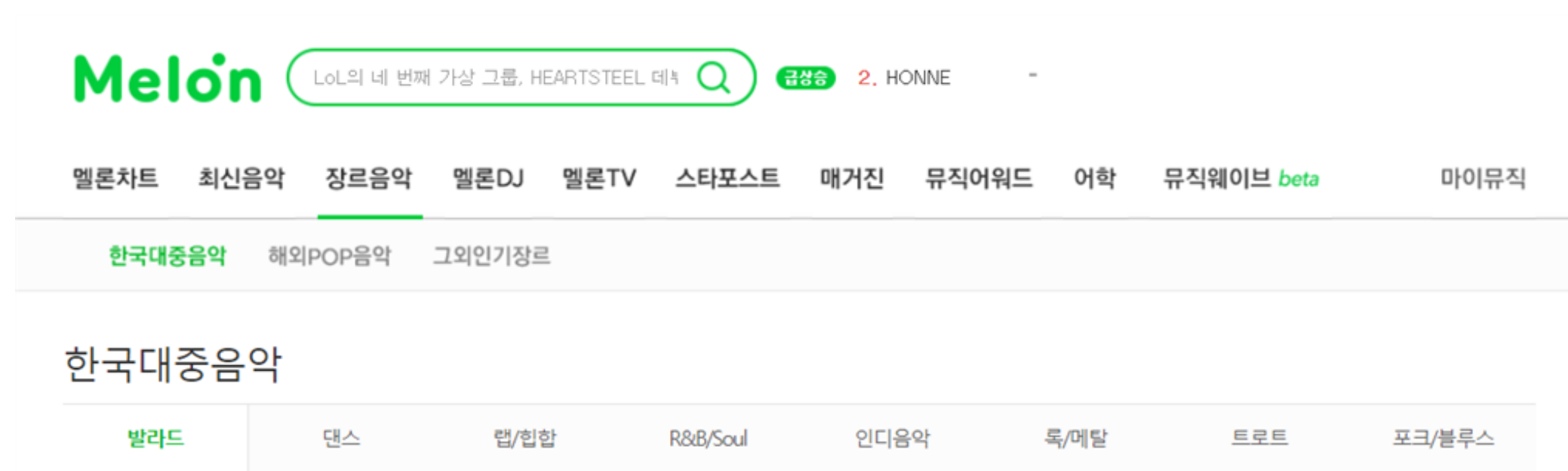
최신순 | 추천순 | 답글순

	현모약처 	박재정의 답이 박재정이었다니.. 2023.04.26 20:43 답글 8개 	 477  8 
	도돌이표 	노래잘하는 잘생긴 발라더인데 여태 안듣게 의문.. 곡운이 없었던건지.. 자작곡으로 이제야 빛을 좀 보나? 2023.04.26 21:44 답글 1개 	 321  4 
	greyhound55 	와 리메이크가 난무하는 요즘 한땀한땀 자작곡으로 발라드차트2위까지 가다니 박재정 대단하다 2023.04.30 14:19 답글 0개 	 319  3 
	lch_ 	요즘 발라드로 10등안에 들어오기 엄청 어려운데 재정님이 해내셨네요 오랫동안 지켜온 팬으로서 너무 기쁘네요 2023.05.10 03:06 답글 2개 	 236  4 

사이트 'MELON'

댓글에는 사람들의 음악 취향과 감정이 담겨있습니다.
이를 바탕으로 맞춤형 플레이리스트 서비스를 제공하고자 합니다.

데 이 터 수 집



- ✓ 한국 대중음악 장르별 + 해외 POP 음악 인기순 150개의 데이터 수집
- ✓ 가수, 노래제목, 앨범사진, 추천순 리뷰 200개 크롤링
- ✓ MariaDB 저장

출처 : 멜론 (Melon)

데이터 수집 전처리

```
# 불용어 정의
stopwords = ['의', '가', '이', '은', '들', '는', '좀', '잘', '강', '과', '도', '를', '으로',
             '자', '에', '와', '한', '하다', '재생', '다운로드', '곡명', '앨범명', '아티스트명', 'π', 'ㅋ', '왁',
             'ㄱㄱ', 'ㄱㄱㄱ', 'ㄱㄱㄹ', 'ㄱㄱㅂ', 'ㄱㄱㅅ', 'ㄱㄴ', 'ㄱㄴㄴ', 'ㄱㄴㄹ', 'ㄱㄴㄹㅇ', 'ㄱㄴㄹㅇㅂ']

# 명사와 형용사만 선택
tokenized_data = []
for s in tqdm(df['총리뷰']):
    sentence = okt.pos(s, norm=True, # norm 정규화
                       stem=True) # 형태소와 품사 정보 추출

# 조사 삭제
selected_words = [word for word, pos in sentence if pos != 'Josa' and word not in stopwords]
tokenized_data.append(" ".join(selected_words))
```

- ✓ 가수, 노래 제목을 primary key로 두어 중복 삭제
- ✓ 댓글 180개 이하 삭제
- ✓ okt로 불용어와 조사 삭제

=> 약 600개 데이터

데 이 터 분 석

```
// = CountVectorizer(  
    ngram_range=(1, 2)  
    min_df=3)
```

‘CountVectorizer’

- ✓ ngram_range: 단일 단어와 두 개의 연속된 단어 조합을 모두 고려
- ✓ min_df: 빈도수 3개 이상

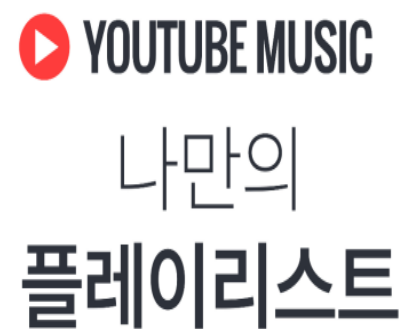
가수	제목	유사도분석
미연 ((여자)아이들)	Drive	0.113307
j-hope	on the street (with J. Cole)	0.038291
지민	Like Crazy	0.031613
태연 (TAEYEON)	Weekend	0.030360

‘코사인 유사도 분석’

- ✓ test = '드라이브'
- ✓ 유사도 높은 순 4개 추출
- ✓ 유사도 분석 값 낮은 이유
 - 데이터 적음
 - 명사, 형용사, 동사 다 사용

HTML 목업

홈 화면



[시작하기] 버튼 클릭 시
입력 화면 이동

입력 화면

홈 버튼

홈 화면으로 이동

검색어 입력

영상이미지

추가 링크 페이지

메인 화면

홈 버튼

홈 화면으로 이동

검색어 입력

앨범 사진

노래 제목

플레이리스트 제목

댓글

추가 링크 페이지

영상이미지

HTML 메인 화면

로고

검색어 입력

앨범 사진

플레이리스트 제목

노래 제목

댓글 등록

music**tb**

ABC **singer**

ABC **song**

ABC **album**

ABC **review1**

ABC **review2**

영상 이미지

play**tb**

ABC **playlist**

ABC **review**

🕒 **datetime**

Description

- ✓ 로고 : 홈 화면으로 이동
- ✓ 검색어 입력 : music**tb** DB 데이터를 가져와 입력 데이터 분석
- ✓ 플레이리스트 제목 : 입력 데이터 출력
- ✓ 앨범 사진, 노래 제목 : 분석 결과 출력
- ✓ 댓글 : 해당 노래의 댓글 각 10개씩 출력 및 play**tb** DB 저장
- ✓ 댓글 등록 : 등록한 댓글을 play**tb** DB에 저장하며 현재 페이지 다시 로드
- ✓ 영상 이미지 : 미리 입력데이터를 지정해둠. 이후 동일

주제

목차

개발환경

주제

데이터수집

데이터분석

HTML

웹시연

웹 시연 마무리

감사합니다