음악서비스 데이터 수집 및 분석

- R을 활용한 가사데이터 분석(feat.Python) -

융복합 프로젝트형 빅데이터 분석 서비스 개발

2020.10.12 발표

박성준

COTENTS

주제선정

자소설닷컴 분석 시도
멜론 플레이리스트 분석

분석 순서

접근 로직
수행결과

개발후기

도전과제

주제선정배경

- 음원 서비스 개선을 지표 전달
- 플레이리스트 내 음원 선정 시 고려할 지표 선정
- 작사가를 위한 특정 주제와 연관이 깊은 단어 정보 제공

핵심부분

플레이리스트 내 가사분석을 통한 연관성 찾기

멜론 곡 정보로 **무엇**을 분석해볼까?

01 | 한국인이 가장 좋아하는 가수는?

02 | 노래 제목 길이와 노래 좋아요 수의 상관관계는?

03 | 플레이리스트 키워드와 가사의 연관성은?

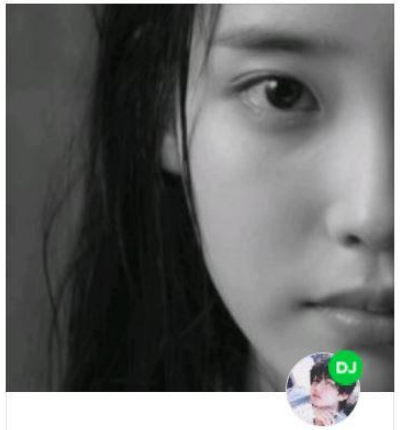
04 | 플레이리스트 좋아요와 플레이리스트 내 노래 좋아요와의 상관관계는?

05 | 노래 가사 전체에서 가장 많이 등장하는 단어는?

06 | 플레이리스트 내 가사 연관분석을 통한 유사성은?

PICTURE

최신곡 | 최신앨범 | 스테디셀러 | 명반 | **추천음악**



왜 눈물이 날까, 이별#2

팔레트속색을섞어

♡ 75 | 37곡

#이별 #슬픔

기분좋은 저녁 산책을 위한 아이돌 발라드

대중가요마스터

♡ 541 | 25곡

#아이돌 #산책

가을 아침, 조용히 귀 기울이는 국내 포크&어쿠스틱

DJ멜군

♡ 2,660 | 45곡

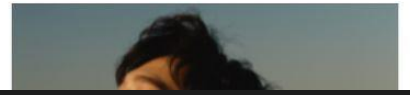
#가을 #아침

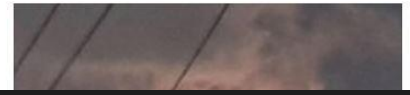
This Week's 취향저격 '발라드' (매주 업데이트)

대중가요마스터

♡ 3,280 | 20곡

#발라드 #취향저격





기준

1. 좋아요 1000개 이상
2. 수록곡 40곡 미만

기준 충족 시 플레이리스트 클릭

PICTURE

1. 플레이리스트 제목 수집
2. 플레이리스트 좋아요 수집

데이터 수집 후 곡 정보 클릭

PICTURE

곡 정보

그대에게

강아솔

앨범 당신이 놓고 왔던 짧은 기억

발매일 2012.04.24

장르 인디음악, 록/메탈

10,452 곡 다운 선물하기

+ 담기

댓글 42개

공유

가사

프린트 가사오류신고

그럴 수없이 사랑하는
나의 벗 그대여
오늘 이 노래로 나 그대를
위로하려하오
하루하루 세상에 짓눌려
얼굴 마주보지 못해도
나 항상 그대 마음

1. 노래 제목 수집
2. 가수 이름 수집
3. 노래 좋아요 수집
4. 가사수집

노래 정보 추출 알고리즘

페이지가 10의 단위에 도달하면 다음 페이지 버튼을 눌러 다음 10개 페이지로 넘어간다.

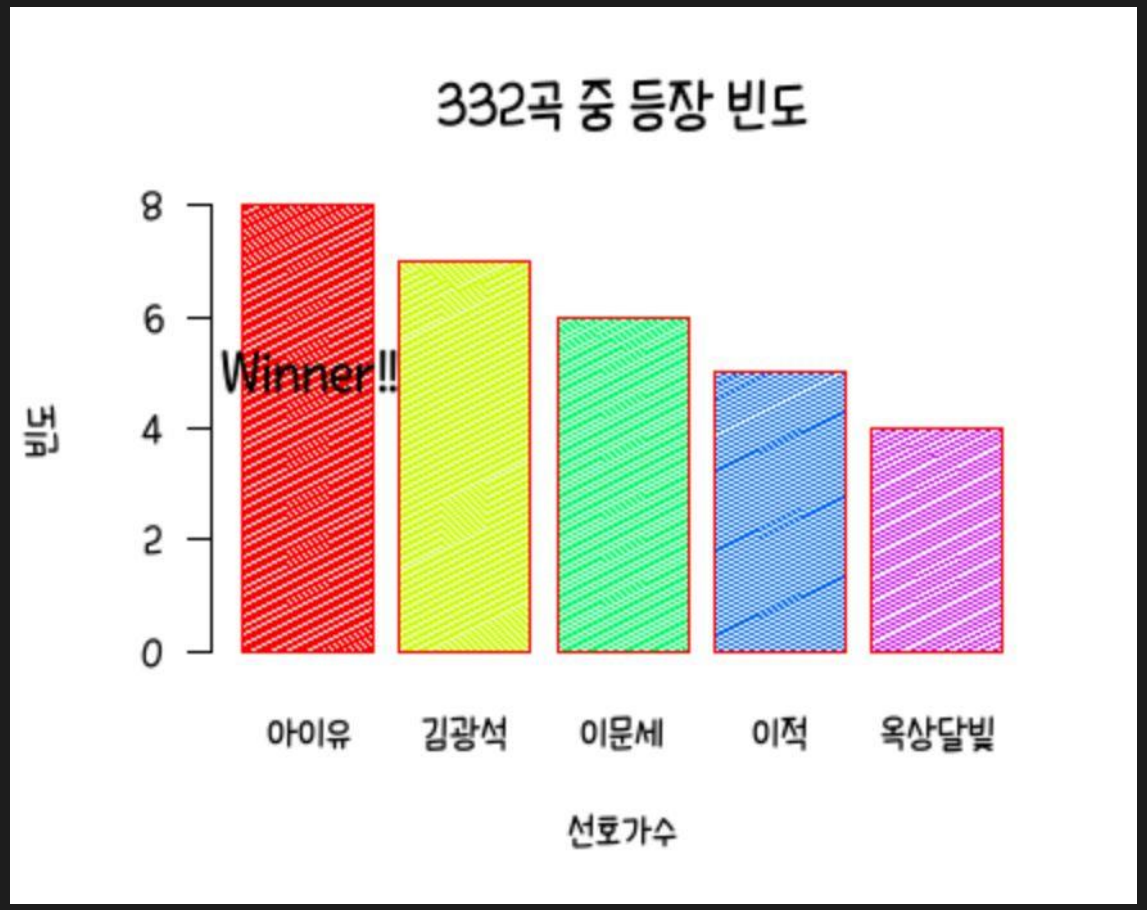
1. 페이지 내 20개 플레이리스트 중 좋아요가 1000개 이상인 플레이리스트를 찾는다.
2. 좋아요가 1000개 이상인 플레이리스트 수록곡이 40개 이하인지 판단한다.
3. 해당 조건이 만족되면 클릭하여 플레이리스 정보로 들어간다.
    - A. 플레이리스트 제목을 추출하여 저장한다.
    - B. 플레이리스트 좋아요 개수를 추출하여 저장한다
    - C. for문을 위한 플레이리스트 내 수록곡 숫자를 추출하여 저장한다.
    - D. 곡 정보를 클릭, 곡 정보 페이지로 이동한다.
        - i. 노래 좋아요 개수를 추출하여 저장한다.
        - ii. 가수 이름을 추출하여 저장한다.
        - iii. 가사를 추출하여 저장한다.
        - iv. 뒤로가기로 돌아가 for문을 수행하기 위한 상태로 돌아간다.
    - E. for문 수행이 끝나면 해당 플레이리스트 제목을 csv파일 제목으로 하여 저장한다.
    - F. 페이지 내 플레이리스트를 찾기 위해 뒤로가기로 돌아간다.
4. 페이지 내 플레이리스트 1000개 이상인 목록이 없을 경우 다음 페이지로 넘어간다.

첫번째 분석

한국인이
가장 좋아하는 가수는?

Barplot을 이용한 시각화

PICTURE



플레이 리스트 내 가장 많이 등장한 가수는 '아이유'

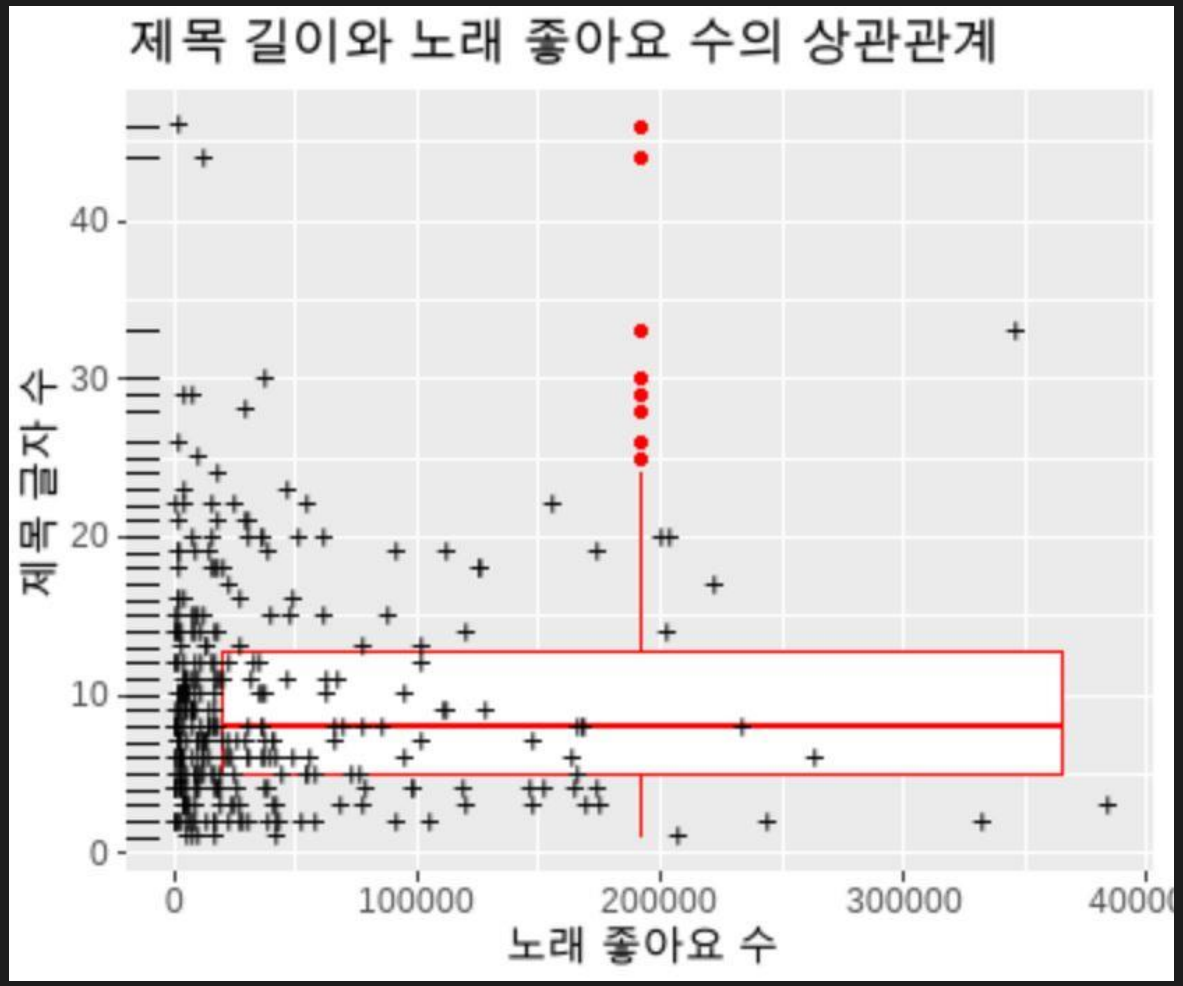
두번째 분석

노래 제목길이의 분포는?

ggplot2를 이용한 시각화

- 1분위 수 : 5글자
- 3분위 수 : 13글자

PICTURE



노래 제목 분포도

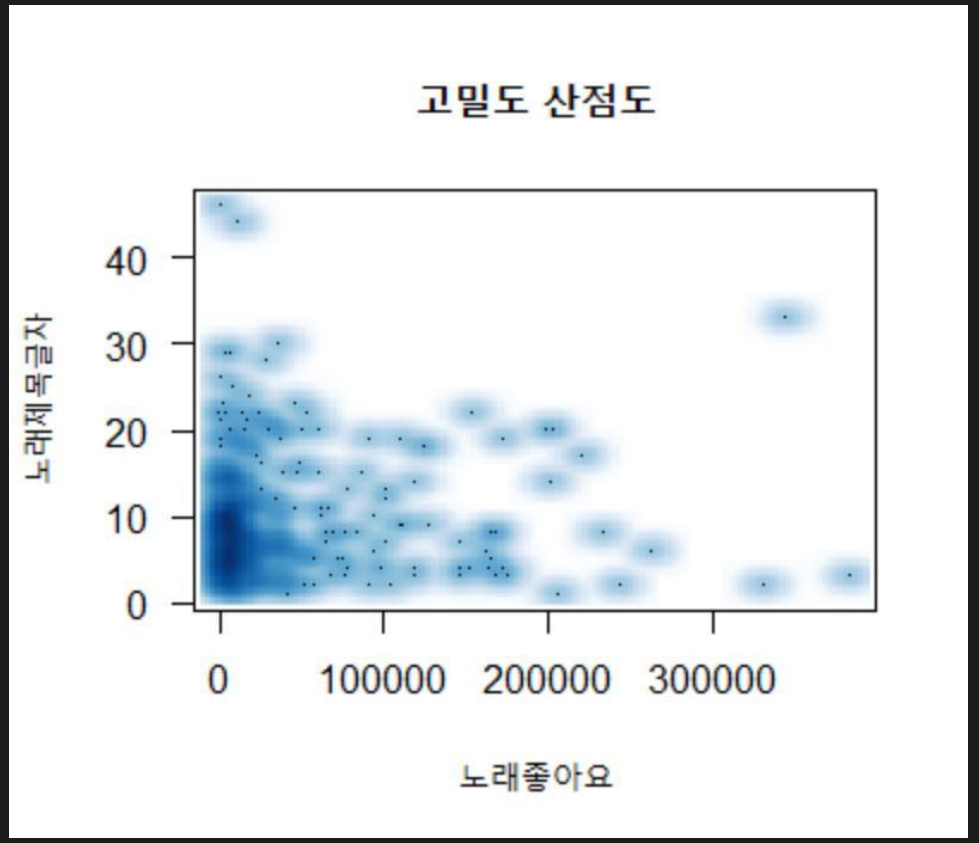
두번째 분석

노래 제목길이와 좋아요 수의 상관관계는?

smoothScatter함수를 이용한 시각화

- 노래 제목 글자 분포 파악

PICTURE



노래 제목과 좋아요 수의 산점도

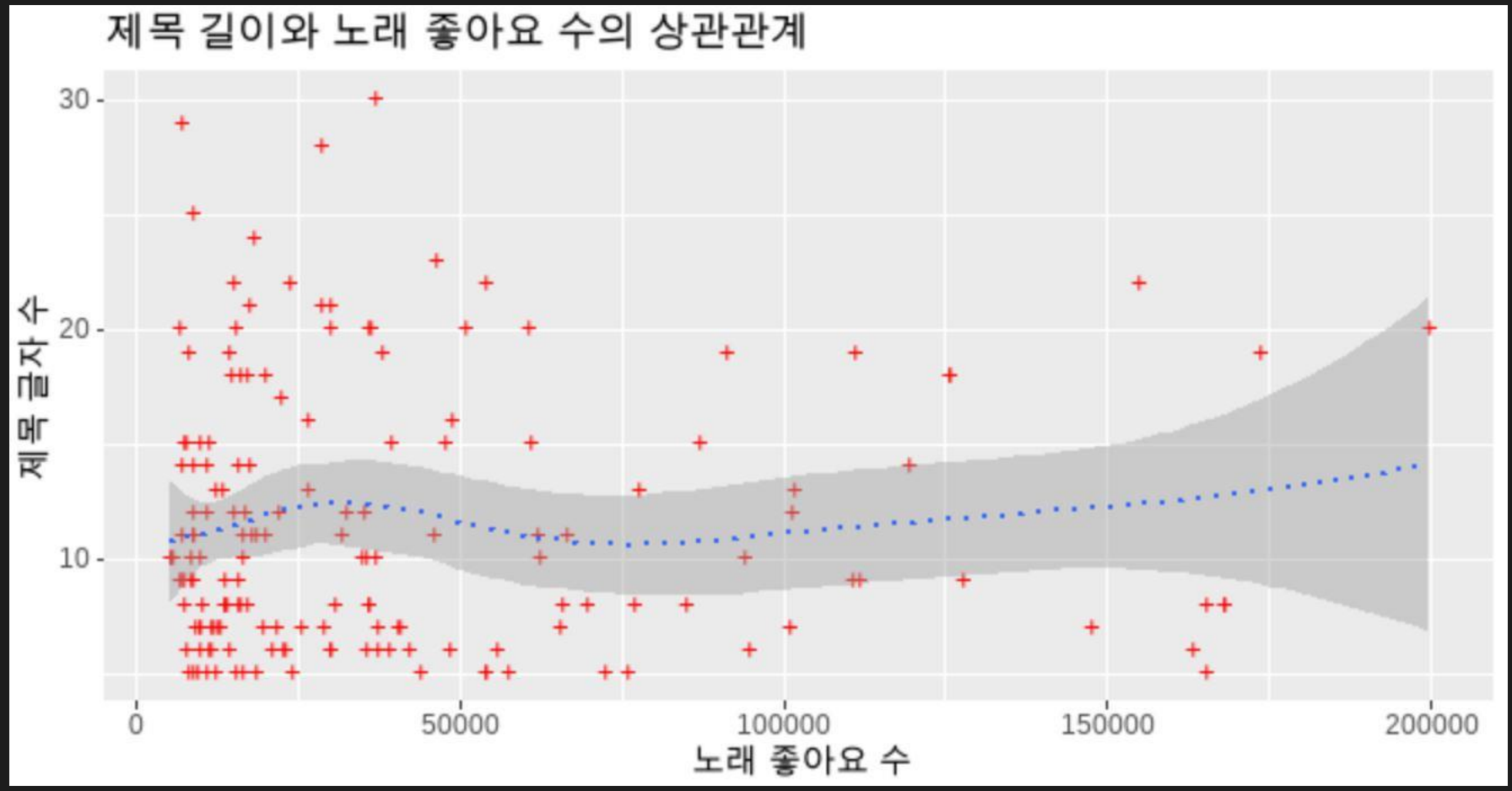
두번째 분석

노래 제목길이와 좋아요 수의 상관관계는?

ggplot2를 이용한 시각화

- 제목 5글자 이상 30글자 이하
- 좋아요 5천 이상 20만 이하

PICTURE



제목 길이와 좋아요 수는 상관관계가 적다

두번째 분석

노래 제목길이와
좋아요 수의 상관관계는?

lm() 함수를 통한 회귀분석

PICTURE

```
> sing_lm <- lm(songlikesum~songtitlenum,data=titlenum)
> summary(sing_lm)

Call:
lm(formula = songlikesum ~ songtitlenum, data = titlenum)

Residuals:
   Min     1Q Median     3Q    Max
-38713 -34183 -22985   2147 345483

Coefficients:
             Estimate Std. Error t value        Pr(>|t|)
(Intercept)  39045.89    5665.87   6.891 0.0000000000294 ***
songtitlenum   -66.75     483.28  -0.138             0.89
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 59090 on 320 degrees of freedom
Multiple R-squared:  5.962e-05,  Adjusted R-squared:  -0.003065
F-statistic: 0.01908 on 1 and 320 DF,  p-value: 0.8902
```

제목 길이와 좋아요 수는 상관관계가 적다
낮은 결정계수, 높은 P-value

세번째 분석

플레이리스트 키워드와 가사의 연관성은?

worddoud2를 이용한 시각화

- 2글자 이상 4글자 이하 명사 추출

- 등장 빈도에 따른 크기 조정

PICTURE

플레이리스트 제목 : 새벽감성 100% 충전 잠 못 이루는 밤에 듣는 발라드

세번째 분석

플레이리스트 키워드와 가사의 연관성은?

worddoud2을 이용한 시각화

- 2글자 이상 4글자 이하 명사 추출

- 등장 빈도에 따른 크기 조정

PICTURE

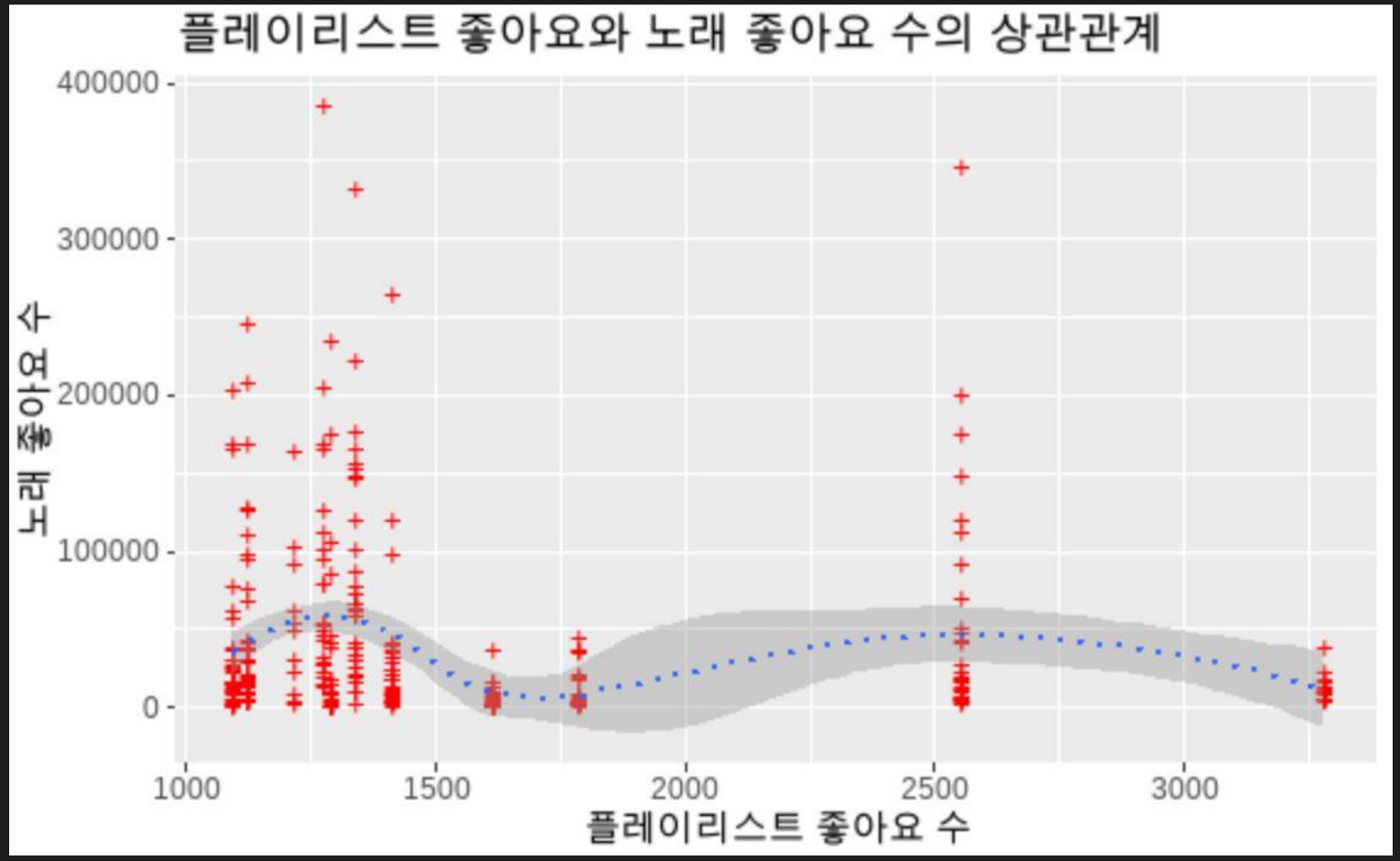
플레이리스트 제목 : 뻔하지 않은 위로곡들과 함께 오늘 하루도 고생했어요

네번째 분석

플레이리스트 좋아요와 노래 좋아요와의 상관관계는?

ggplot2를 이용한 시각화

PICTURE



플레이리스트 좋아요와 노래좋아요 수는 상관관계가 적다

네번째 분석

플레이리스트 좋아요와
노래 좋아요와의
상관관계는?

lm() 함수를 통한 회귀분석

PICTURE

```
> like_lm <- lm(songlikesum~playlistlikesum, data=infoSum)
> summary(like_lm) # 별로 크게 상관없다

Call:
lm(formula = songlikesum ~ playlistlikesum, data = infoSum)

Residuals:
   Min     1Q Median     3Q    Max
-43926 -34245 -21438    675 342352

Coefficients:
                Estimate Std. Error t value      Pr(>|t|)
(Intercept)     56646.811   9239.704   6.131 0.00000000257 ***
playlistlikesum   -11.505      5.452  -2.110        0.0356 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 58690 on 320 degrees of freedom
Multiple R-squared:  0.01373,   Adjusted R-squared:  0.01065
F-statistic: 4.454 on 1 and 320 DF,  p-value: 0.03559
```

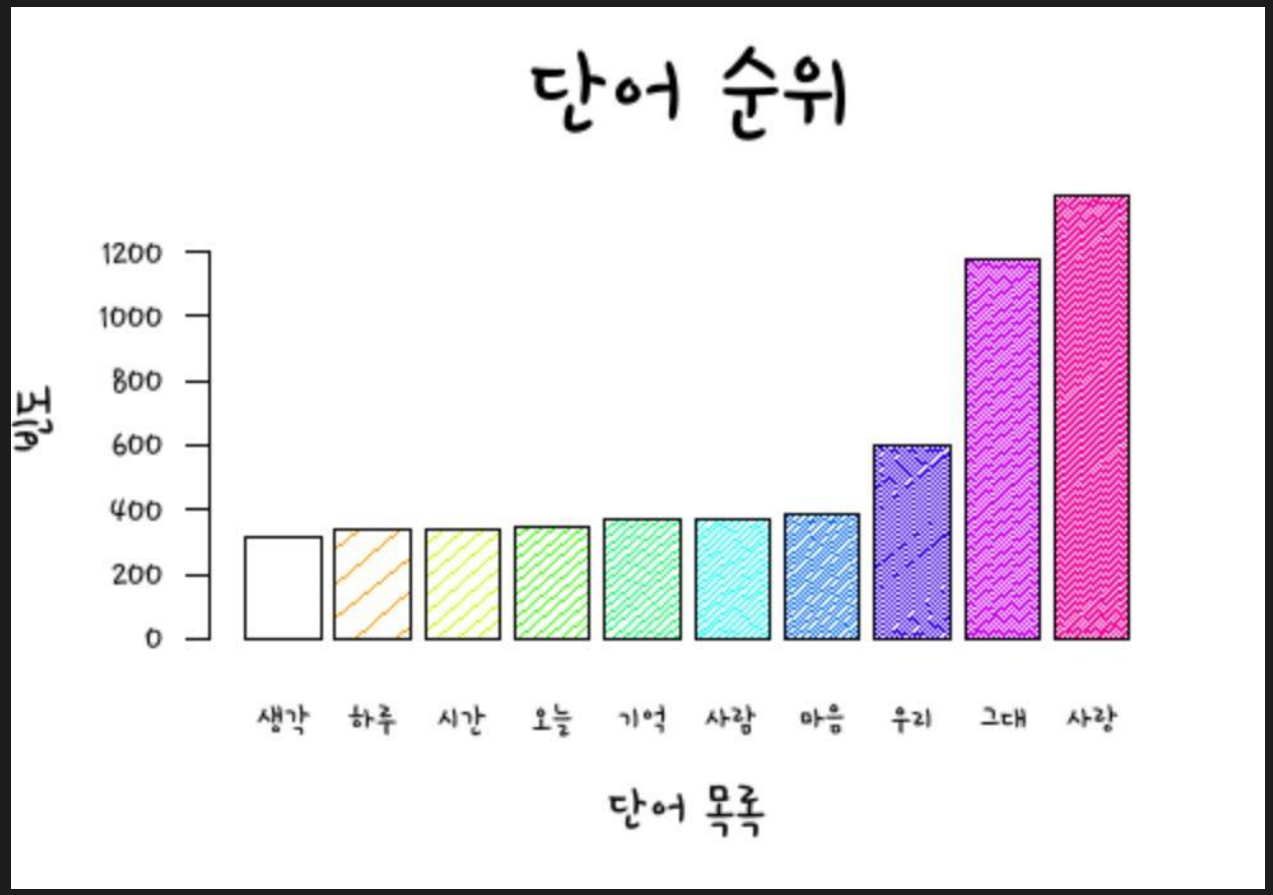
제목 길이와 좋아요 수는 상관관계가 적다
낮은 결정계수, 낮은 P-value

다섯번째 분석

전체 노래 가사에서
가장 많이 등장한 단어는?

barplot을 이용한 시각화

PICTURE



가장 많이 등장한 단어는 '사랑'

여섯번째 분석

플레이리스트별 가사 단어들의 연관분석

apriori를 이용한 시각화

PICTURE

```
> wordtran <- as(strsplit(lyric_filter, " "), "transactions")
> wordtran
transactions in sparse format with
 1452 transactions (rows) and
 1452 items (columns)
> tranrules <- apriori(wordtran, parameter = list(supp =0.01, conf = 0.01))
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen maxlen target  ext
       0.01    0.1    1 none FALSE            TRUE       5    0.01      1     10  rules TRUE

Algorithmic control:
 filter tree heap memopt load sort verbose
    0.1 TRUE TRUE  FALSE TRUE    2    TRUE

Absolute minimum support count: 14

set item appearances ...[0 item(s)] done [0.00s].
set transactions ...[1452 item(s), 1452 transaction(s)] done [0.00s].
sorting and recoding items ... [0 item(s)] done [0.00s].
creating transaction tree ... done [0.00s].
checking subsets of size 1 done [0.00s].
writing ... [0 rule(s)] done [0.00s].
creating S4 object  ... done [0.00s].
```

지지도와 신뢰도를 낮게 설정해도 출력X
안되면 Python 으로 시도해보자!!

여섯번째 분석

플레이리스트별 가사 단어들의 연관분석

apriori를 이용한 시각화

PICTURE

	support	itemsets
0	0.657895	(그대)
1	0.631579	(사랑)

	support	itemsets
1	0.657895	(그대)
8	0.631579	(사랑)
13	0.473684	(사랑, 그대)
2	0.315789	(기억)
10	0.315789	(하늘)
0	0.289474	(가슴)
6	0.289474	(바람)
5	0.263158	(마음)
11	0.236842	(가슴, 사랑)
16	0.236842	(사랑, 마음)

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(가슴)	(사랑)	0.289474	0.631579	0.236842	0.818182	1.295455	0.054017	2.026316
1	(사랑)	(가슴)	0.631579	0.289474	0.236842	0.375000	1.295455	0.054017	1.136842
2	(마음)	(그대)	0.263158	0.657895	0.236842	0.900000	1.368000	0.063712	3.421053
3	(그대)	(마음)	0.657895	0.263158	0.236842	0.360000	1.368000	0.063712	1.151316
4	(사랑)	(그대)	0.631579	0.657895	0.473684	0.750000	1.140000	0.058172	1.368421
5	(그대)	(사랑)	0.657895	0.631579	0.473684	0.720000	1.140000	0.058172	1.315789
6	(그대)	(하늘)	0.657895	0.315789	0.210526	0.320000	1.013333	0.002770	1.006192
7	(하늘)	(그대)	0.315789	0.657895	0.210526	0.666667	1.013333	0.002770	1.026316
8	(기억)	(사랑)	0.315789	0.631579	0.210526	0.666667	1.055556	0.011080	1.105263
9	(사랑)	(기억)	0.631579	0.315789	0.210526	0.333333	1.055556	0.011080	1.026316

플레이리스트 제목 : 가을바람 생각나는 7080 추억의노래
플레이리스트와 가장 연관이 깊은 단어는 '그대' 와 '사랑'

여섯번째 분석

플레이리스트별
연관분석을 통해
비슷한 주제 찾기

apriori를 이용한 시각화

PICTURE

	support	itemsets
0	0.875	(사랑)

	support	itemsets
7	0.87500	(사랑)
4	0.46875	(다시)
2	0.43750	(기억)
21	0.43750	(사랑, 다시)
3	0.40625	(눈물)
6	0.40625	(사람)
18	0.40625	(기억, 사랑)
11	0.40625	(이별)
23	0.34375	(사랑, 사람)
20	0.34375	(눈물, 사랑)

폭풍눈물주의
절절한 이별 노래

	support	itemsets
0	0.533333	(사랑)

	support	itemsets
10	0.533333	(사랑)
16	0.366667	(이별)
2	0.333333	(기억)
5	0.300000	(다시)
6	0.300000	(마음)
7	0.300000	(모든)
14	0.266667	(오늘)
18	0.266667	(정말)
15	0.233333	(우린)
26	0.233333	(이별, 사랑)

새벽감성 100% 충전
잠 못 이루는 밤에
듣는 발라드

개발 후기

- 연관성 규칙에 대해 좀 더 공부해보고 싶다!!

도전하고 싶은 것

음파를 수치화해서
플레이리스트 비교 분석?

음악서비스 데이터 수집 및 분석

- R을 활용한 가사데이터 분석(feat.Python) -

Thank you