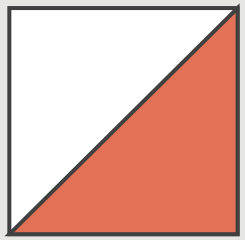


현대대중가요 가사 분석

172114 임성애

목차

1. 프로젝트 개요

현대대중가요 **장르별&실시간** 분석을 통해

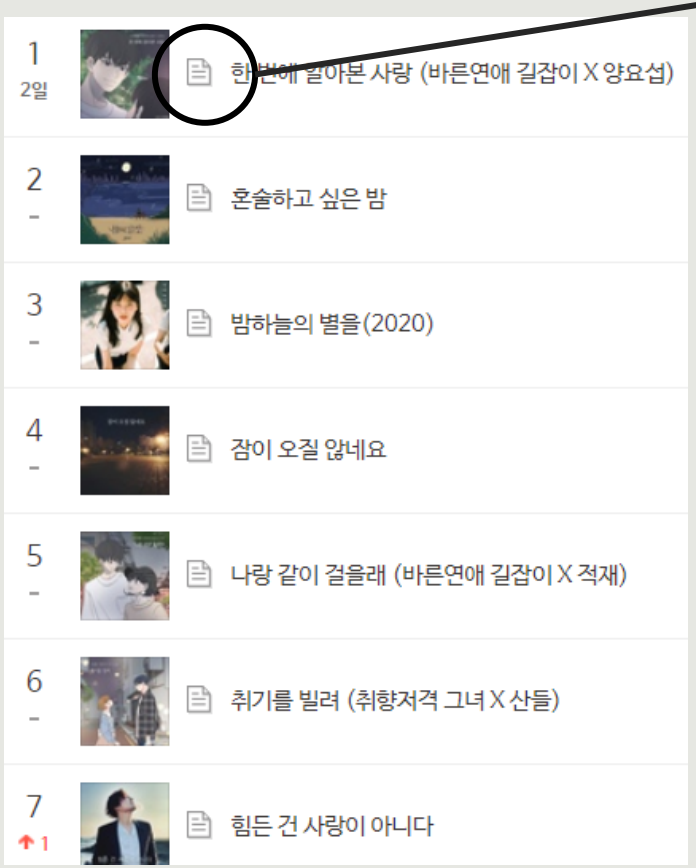
대중의 생각과 관심을 **빠르게 파악**

빅데이터 시스템을 이용한

국내 음원 '장르별 Top 100 가사' 분석

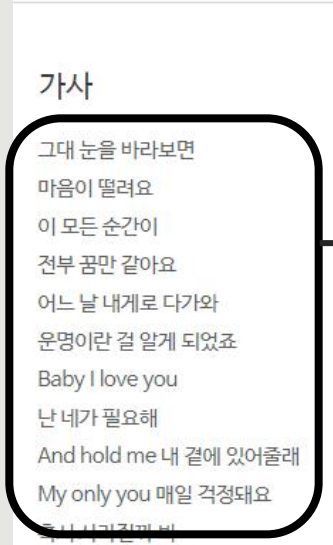
2. 프로젝트 진행

Top 100 가사 가져오기



ex. 발라드 Top 100

```
<a href="https://music.bugs.co.kr/track/6028763?wl_ref=list_tr_08_chart" class="trackInfo" onclick="
">곡정보</a>
```




```
▼<div class="lyricsContainer">
  <p></p>
  ▼<xmp>
    "그대 눈을 바라보면
    마음이 떨려요
    이 모든 순간이
    전부 꿈만 같아요
    어느 날 내게로 다가와
    운명이란 걸 알게 되었죠
    Baby I love you
    난 네가 필요해
    And hold me 내 곁에 있어줄래
    My only you 매일 걱정돼요
    혹시 사라질까 봐
```

2. 프로젝트 진행

Top 100
가사
가져오기

['그대 눈을 바라보면\r\n마음이 떨려요\r\n이 모든 순간이\r\n전부 꿈만 같아요\r\n어느 날 내게로 다가와\r\n운명이란 걸 알게 되었죠\r\nBaby I love you\r\n난 네가 필요해\r\nAnd hold me 내 곁에 있어줄래\r\nMy only you 매일 걱정돼요\r\n혹시 사라질까 봐\r\n그댄 나의 사랑\r\nOnly you\r\n내게 보였던 그 미소\r\n다 너무 예뻐서\r\n더 보고 싶어서\r\n나를 달래고 있죠\r\n한 번에 알아 본 사랑은\r\n내 생에 오직 그대뿐이죠\r\nBaby I love you\r\n난 네가 필요해\r\nAnd hold me 내 곁에 있어줄래\r\nMy only you 매일 걱정돼요\r\n혹시 사라질까 봐\r\n그댄 나의 사랑\r\nOnly you\r\n푸르고 짙은 계절이 찾아올 때\r\n그대 나의 꽃이 되어 주기를\r\nIn my heart\r\nBaby I love you\r\n나 너를 사랑해\r\nAlways love you 곁에 있어 줄래\r\nMy only you 매일 걱정돼요\r\n혹시 사라질까 봐\r\n그댄 나의 사랑\r\nOnly you', '위로받고 싶은 날 내겐 말해도 돼\r\n위로받고 싶은 날 내겐 기대도 돼\r\n비에 섞인 음악 소리에\r\n혼자 젖은 새벽 감성에\r\n때론 너무 지치고 때론 너무 지쳐서\r\n가끔 아주 가끔\r\n혼술하고 싶은 밤 전화해\r\n혼자 있고 싶은 밤 전화해\r\n참았던 눈물에 울컥해도 괜찮아\r\n가끔 아주 가끔\r\n혼술하고 싶은 밤 전화해\r\n잠이 오지 않는 이 밤에\r\n한 번쯤 소리 내 울컥해도 괜찮아\r\n가끔 아주 가끔\r\n웃고 싶지 않은 날 그냥 그래도 돼\r\n도망치고 싶은 날 그래 떠나도 돼\r\n비에 섞인 눈물 소리에\r\n젖어 드는 슬픈 감정에\r\n때론 너무 지치고 때론 너무 지쳐서\r\n가끔 아주 가끔\r\n혼술하고 싶은 밤 전화해\r\n혼자 있고 싶은 밤 전화해\r\n참았던 눈물에 울컥해도 괜찮아\r\n가끔 아주 가끔\r\n혼술하고 싶은 밤 전화해\r\n잠이 오지 않는 이 밤에\r\n한 번쯤 소리 내 울컥해도 괜찮아\r\n가끔 아주 가끔\r\n전부 이해할 수 없지만 내게 말해\r\n참았던 눈물이 왈칵 쏟아질까 봐\r\n미치게 답답해 울컥하는 마음에\r\n가끔 기대도 돼 내게\r\n혼술하고 싶은 밤 외로워\r\n혼자라는 생각

ex. 발라드 Top 100 가사 전체

2. 프로젝트 진행

가사 전처리

KoNLPy

Kkma	Hunnanum	Okt
많고 많은 사람 중에 너를 만나서	많고 많은 사람 중에 너를 만나서	많고 많은 사람 중에 너를 만나서
행복하고 싶어 두 번 다시 울지 않을래	행복하고 싶어 두 번 다시 울지 않을래	행복하고 싶어 두 번 다시 울지 않을래
오직 내 눈에는 너만 보여 나를 아껴줘	오직 내 눈에는 너만 보여 나를 아껴줘	오직 내 눈에는 너만 보여 나를 아껴줘
이제부터 혼자가 아니야 우린 함께니까	이제부터 혼자가 아니야 우린 함께니까	이제부터 혼자가 아니야 우린 함께니까
나나나 난난 난난나	나나나 난난 난난나	나나나 난난 난난나
나나나 난난 난난나	나나나 난난 난난나	나나나 난난 난난나
나 나나나 난난 난난나 나나난나	나 나나나 난난 난난나 나나난나	나 나나나 난난 난난나 나나난나

ex. 발라드 Top 100 가사 단어

```
['그대', '눈', '마음', '순간', '전부', '꿈', '날', '운명', '필요', '내', '곁', '
'걱정', '나의', '사랑', '미소', '나', '번', '생', '계절', '때', '꽃', '주기', '
'너', '위로', '날', '내', '내겐', '겐', '기대', '비', '음악', '소리', '혼자', '
'젖', '새벽', '감성', '때', '혼', '혼술', '술', '밤', '전화', '눈물', '잠', '번
', '번쯤', '쯤', '내', '감정', '이해', '수', '마음', '워', '생각', '방하늘', '
'별', '너', '내가', '사랑', '내', '곁', '어색', '널', '처음', '날', '사람', '숨
', '말', '손', '밤', '하늘', '마음', '표현', '꿈', '웃음', '대도', '겐', '순간
', '영원', '행복', '중', '번', '눈', '나', '이제', '혼자', '니', '난', '당신',
 '날', '내', '마음', '밤', '잠', '매일', '생각', '유치', '심장', '춤', '나', '
'너', '워', '하나', '그대', '오늘', '아름', '그', '고요', '워오', '오', '가을',
'가을밤', '밤', '그대', '나', '눈', '때', '향기', '맘', '지네', '너', '생각', '
'별', '마음', '순간', '내일', '부담', '날', '산책', '너와', '딴', '보고', '널',
```

2. 프로젝트 진행

하둡

데이터 크기↑

id	word
5366	그림
5367	남아
5368	주기
5369	나
5370	넌
5371	내
5372	지난
5373	시간
5374	햇살
5375	너
5376	손위
5377	보석
5378	영원
5379	약속

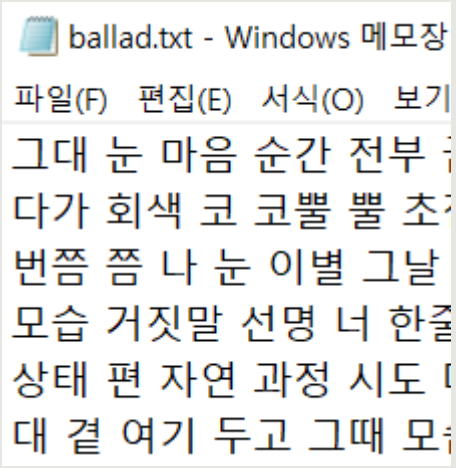
한글 깨짐 현상

2. 프로젝트 진행

하둡

Maven™ eclipse

메모장 저장



Pom.xml에서 하둡 라이브러리 넣기

```
<dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-mapreduce-client-core</artifactId>
    <version>3.1.4</version>
</dependency>
<dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-common</artifactId>
    <version>3.1.4</version>
</dependency>
<dependency>
    <groupId>org.apache.hadoop</groupId>
    <artifactId>hadoop-mapreduce-client-jobclient</artifactId>
    <version>3.1.4</version>
</dependency>
```

기존 wordcount.java 이용

```
public static class TokenizerMapper
public static class IntSumReducer
```

[빈도수 2이상 단어 추출] 코드

2. 프로젝트 진행

하듑

가슴속	3
가요	5
가운데	2
가을	3
가을밤	2
가지	4
갈게	3
감	3
감정	4

발라드 Top 100 단어 맵리듀스

가요	2
가운데	2
가지	3
가치	2
간직	2
갈게	3
갈래	4
감	7
감사	2

댄스

가요	3
가자	3
가족	2
가지	4
각자	4
간	6
간지	2
갈게	6

힙합

가도	2
가슴	27
가슴속	5
가요	5
가운	2
가지	6
가지마	2
간	4
간직	5

OST

거기	2
거리	11
거울	2
거짓	2
거짓말	3
걱정	6
건	8
걸	3
게	4

어쿠스틱

가까이	2
가로등	3
가사	4
가슴	13
가야	3
가요	4
가을	2
가지	5
간	2

알앤비

2. 프로젝트 진행

시각화

python 워드 클라우드(wordcloud)

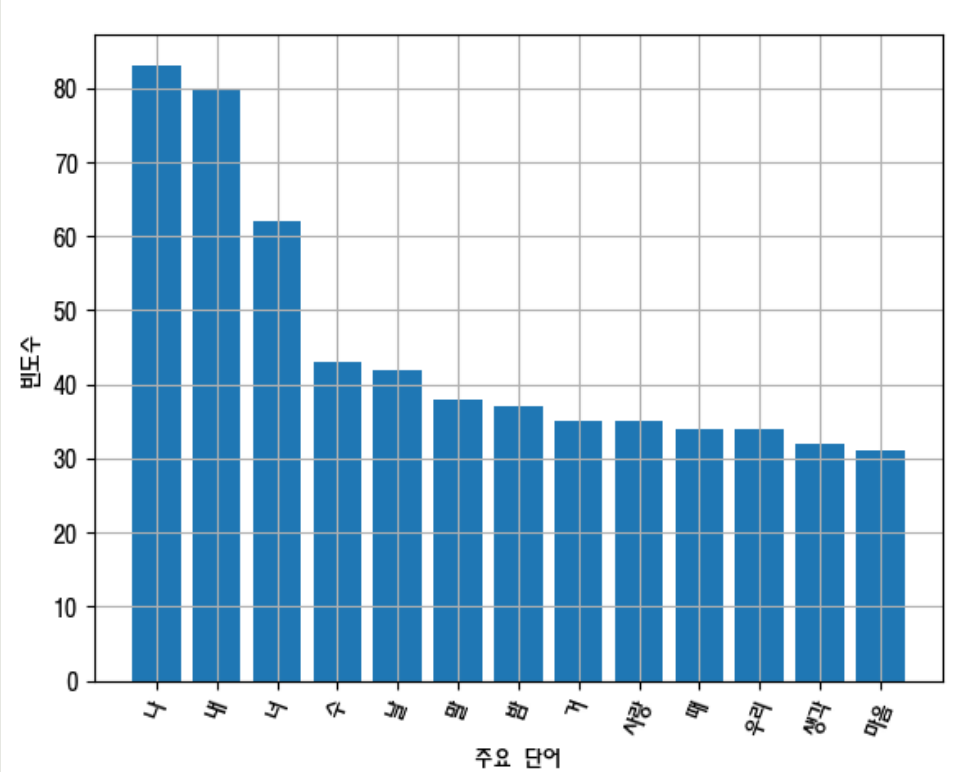
발라드

힙합

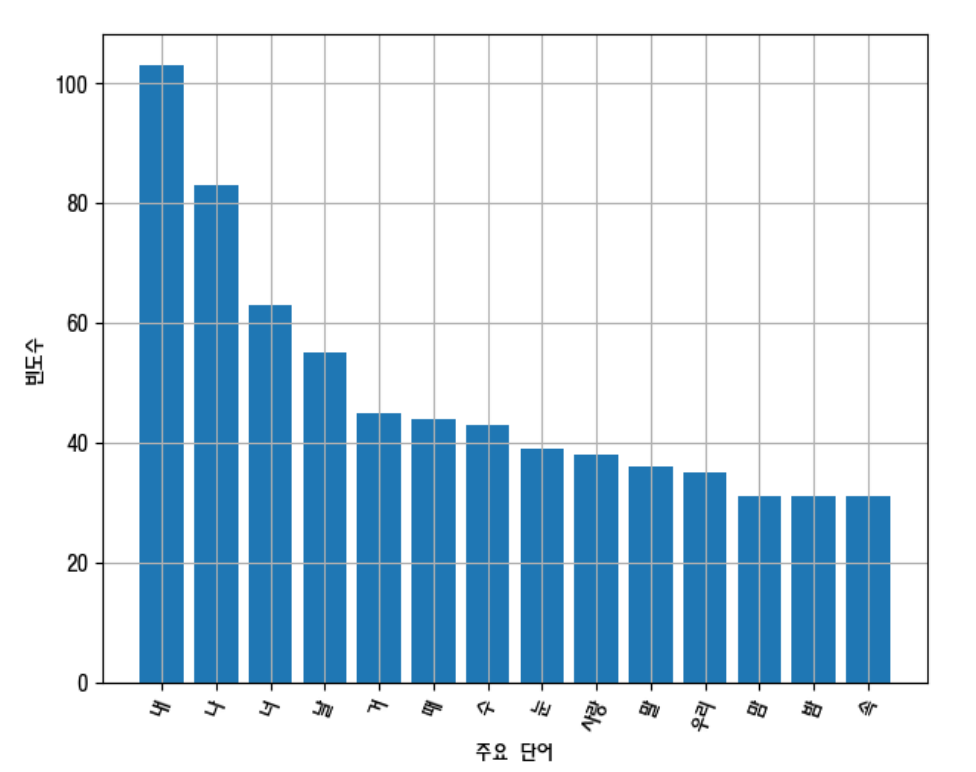
2. 프로젝트 진행

시각화

python 그래프(matplotlib)



어쿠스틱



댄스

2. 프로젝트 진행

시각화

토픽 모델링(pyLDAvis)

여러 개의 문서의 공통된 주제를 찾아내는 프로세스

발라드

댄스

힙합

OST

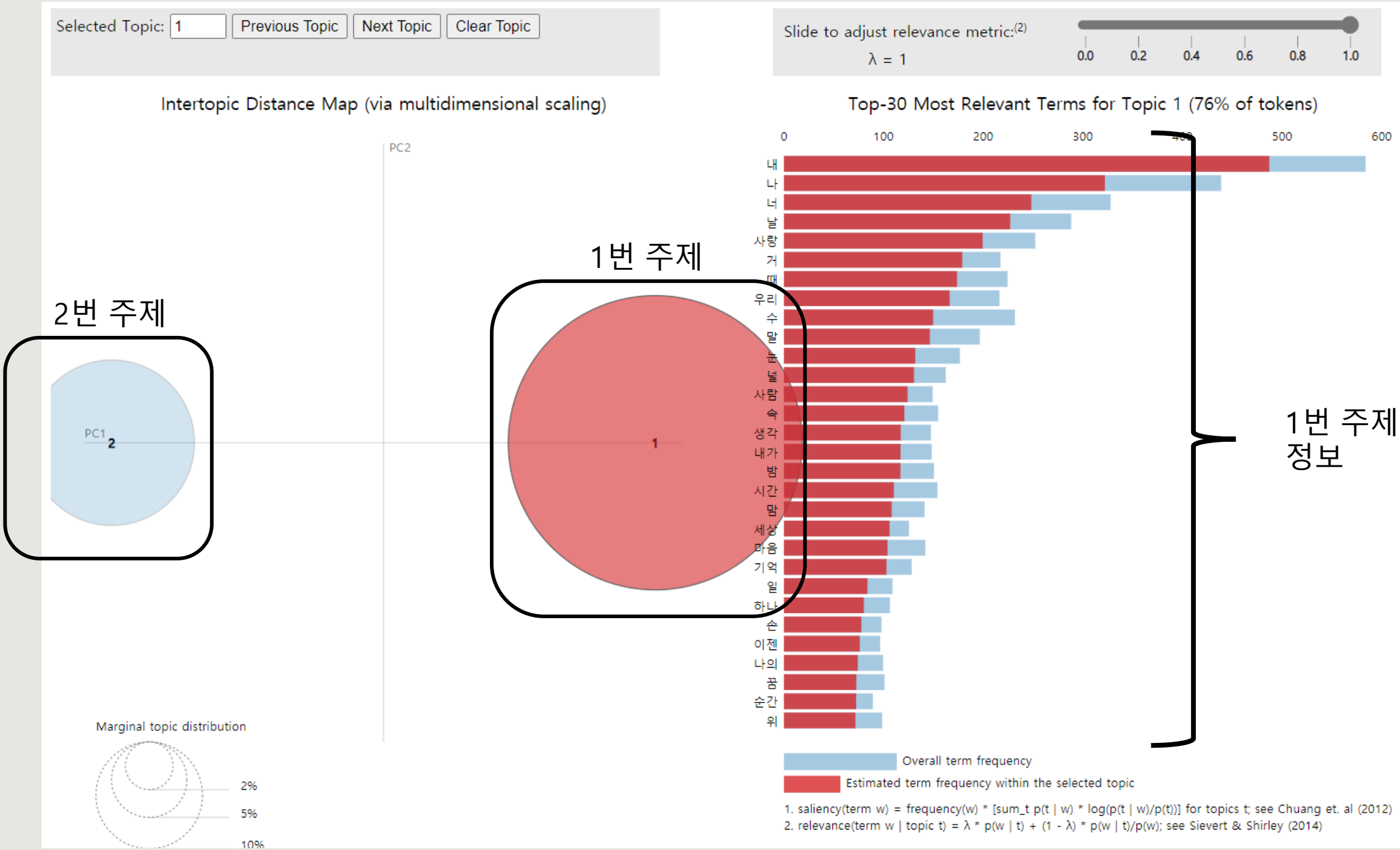
어쿠스틱

알앤비

Input data

-단어 정수 인코딩
-단어 빈도수

→ **공통된 주제**


Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
2
1
2번 주제
1번 주제
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (76% of tokens)
0
100
200
300
500
600
내
나
너
날
사랑
거
때
우리
수
말
눈
널
사람
속
생각
내가
밤
시간
맘
세상
마음
기억
일
하나
손
이젠
나의
꿈
순간
위
1번 주제 정보
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

4. 프로젝트 결론

기대효과

-기존 가사 분석보다 실시간으로 다양하게 분석 가능

-대중들의 관심과 트렌드를 빠르게 파악 가능

-나의 음원 목록 분석 등 다른 분야 분석 가능

한계점

-한글 명사 추출이 깔끔하게 되지 않음 (ex. 갈 수 있다 –> ‘갈’, ‘수’ 추출)

-장르별로 빈도수 높은 단어들이 비슷함