



2009.01 – 2018.11

**음악 순위 예측 및
시대별 트렌드 변화 시각화**

INDEX

- 01 프로젝트 개요
- 02 데이터 정보
- 03 크롤링 및 데이터 전처리
- 04 분류 모델 생성 및 성능
- 05 Django 연동



프로젝트 개요



//

가수, 작곡가, 장르만으로

//

TOP 10이 가능할 지 예측할 수 있을까?

음악 산업의 발달로 언제 어디서나 음악을 즐길 수 있게 되었다.
가수, 작곡, 장르 데이터를 기반으로 발매할 음악의 순위를 예측하고
지난 10년 간 음악 산업의 흐름과 변화를 알아보자

데이터 정보

엠넷 뮤직 월간 종합 차트

데이터 기간: 2009.01 – 2018.11, 총 10년 간

데이터 칼럼: 1위부터 100위까지 순위정보,
곡명, 가수, 작곡가, 작사가, 피쳐링, 발매일,
장르, 재생시간, 좋아요, 편곡, 활동유형

엠넷 뮤직 선정 이유

월 별 차트 접근 시 곡 데이터 정보를 수집하기 용이하도록
url 및 tag 정리가 잘 되어 있음

Mnet



TV

뮤직

차트

최신음악

장르음악

플레이리스트

뮤직비디오

매거진

종합 차트

장르별 차트

아티스트 차트

앨범 차트

엠카운트다운 차트

실시간 | 일간 | 주간 | 월간 | 연간

< 2018.11 >

1개월간의 물기 40% 다른 60% 집계

전체보기 | 보기 | 다운로드 | 담기

순위 | 등락 | 곡정보 /

1 NEW

2 NEW

3 ↑ 41
PEAK MONTH 2

4 NEW

5 ↑ 12
PEAK MONTH 2

6 NEW

7 ↓ 5
PEAK MONTH 2

8 NEW

1월	2월	3월	4월
5월	6월	7월	8월
9월	10월	11월	12월

너는 어땠을까
노들 / 뽕

내 생애 아들다운
케이윌 / 뷰티 인사이드 OST Part 4

아름답고도 아프구나
비 / 루비 / HOUR MOMENT

뽕뽕
아이유 (IU) / 뽕뽕

수퍼비와 (Prod. BewhY) (Feat. BewhY)
SUPERBEE / 쇼미더머니 777 Semi Final

추천 달기 영상 다운로드 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

추천 + | | MP3 원음

웹 크롤링 방법

1. 월 정보 가져오기

url: <http://www.mnet.com/chart/TOP100/201811>

월간차트 링크 만들기

```
chart_links = []
month = ['01', '02', '03', '04', '05', '06', '07', '08', '09', '10', '11', '12']
for year in range(2009, 2019):
    for m in month:
        # print(str(year) + m)
        chart_links.append('http://www.mnet.com/chart/TOP100/' + str(year) + m)
```

2. 곡 상세정보 가져오기

"201810_4": {"곡명": "하루도 그대를 사랑하지 않은 적이 없었다", ... }

•
•
•

12,000개의 데이터를 dict type으로 MongoDB에 저장





실시간 | 일간 | 주간 | 월간 | 연간

< 2018.10 >

1개월간의 듣기 40% 다운 60% 집계 >

전체듣기 듣기 다운로드 담기 Now Playing 추가



순위	등급	곡정보/곡명	추천	담기	영상	다운로드	원음
1	↑ 57 PEAK 1 MONTH 2	 가을 떠나 봐 바미브 / 가을 떠나 봐	추천	+	화	MP3	원음
2	NEW	 뽀빠 아이유(IU) / 뽀빠	추천	+	화	MP3	원음
3	NEW	 이별길 (GOODBYE ROAD) IKON / NEW KIDS : THE FINAL	추천	+	화	MP3	원음
4	↑ 1 PEAK 4 MONTH 2	 하루도 그대를 사랑하지 않은 적이 없었다 임창정 / 하루도 그대를 사랑하지 않은 적이 없었다	추천	+	화	MP3	원음

월별 메인 페이지에서 순위, 곡별 상세 페이지링크, 곡명 수집

하루도 그대를 사랑하지 않은 적이 없었다

Today 13 48 듣게

곡 소개



하루도 그대를 사랑하지 않은 적이 없었다 (04:03)

임창정

발매일 2018.09.19

음악장르 가요 > 발라드

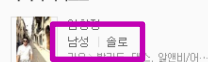
작사 임창정, 멧돼지, 신형섭
작곡 임창정, 멧돼지, 신형섭
더 많은 참여스텝 보기 >

♡ 좋아요 <2,923

참여스텝

보컬	임창정
작사	임창정
작곡	임창정, 멧돼지, 신형섭
편곡	임창정, 멧돼지

이 곡의 아티스트



이 곡이 수록된 앨범



미나차트

상세 페이지에서 재생시간, 발매일, 음악장르, 좋아요, 참여스텝, 활동유형 정보 수집

데이터 전처리

1. 곡명에는 피쳐링 정보(feat.____)가 있으나 피쳐링 칼럼 안에는 정보가 없는 경우

곡명에서 Feat.이 들어간 index를 찾아, 다음 글자부터 ')' 앞까지의 글자를 추출 후 피쳐링 칼럼에 삽입

	곡명	피쳐링
0	Gee	없음
1	이젠 남이야 (Feat. Baby-J of Jewelry)	없음
2	Strong Baby (승리 solo) 19	없음
3	러브119 (Love119) (Feat. MC몽)	MC몽



	곡명	피쳐링
0	Gee	없음
1	이젠 남이야 (Feat. Baby-J of Jewelry)	Baby-J
2	Strong Baby (승리 solo) 19	없음
3	러브119 (Love119) (Feat. MC몽)	MC몽

2. 가수, 작곡가, 작사가가 복수인 경우

'/' 로 구분되어 있는 n명의 사람들을 n개의 칼럼으로 분리

	작곡	작곡1	작곡2	작곡3	작곡4	작곡5	작곡6
11745	Pdogg/Ray Michael Djan Jr/Ashton Foster/Lauren...	Pdogg	Ray Michael Djan Jr	Ashton Foster	Lauren Dyson	RM	정바비
11746	Slom/자이언티/오혁	Slom	자이언티	오혁	없음	없음	없음

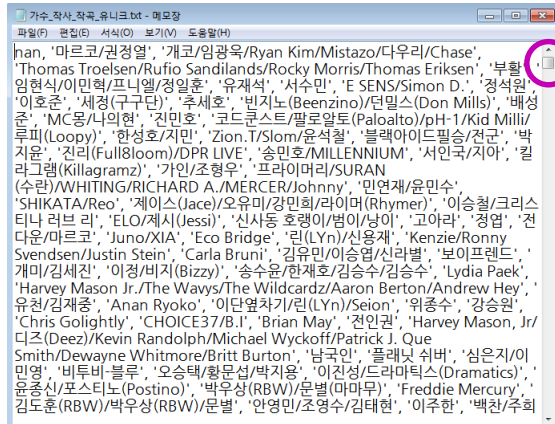
전처리 중 고난과 역경(1)

가수, 작사, 작곡 중복 데이터

데이터 전 처리 과정에서 가수, 작곡가, 작사가 칼럼에서 같은 사람이 다른 이름으로 표현되는 경우

ex. 다이나믹 듀오 ,Dynamic Duo, 다이나믹 듀오(Dynamic Duo)

해결방안



가수, 작사가, 작곡가 unique data 5,852개

mnet_music ☆

파일 수정 보기 삽입 서식 데이터 도구 부가기능 도움말 3일 전에 마지막으로 수정했습니다.

100% W % .0 .00 123 Arial 11 B I U A

	A	B	C	D	E	F	G
		순위	곡명	가수	작곡	작사	피쳐링
1	0	1위	Gee	소녀시대	이트라이브	이트라이브	nan
2	1	2위	이젠 남이야 (Feat. Baby-J of Jewell)	김경록	조영수	조영수	Baby-J
3	2	3위	Strong Baby (송리 solo) 19	송리	배진렬/지드래곤	배진렬/지드래곤	nan
4	3	4위	러브119 (Love119) (Feat. MC몽)	케이윌	조영수/오성훈	장근이/MC몽	MC몽
5	4	5위	U R Man	SS501	한상원	한상원	nan
6	5	6위	Stylish (The FILA)	빅뱅	페리	지드래곤	nan
7	6	7위	내 머리가 나빠서	SS501	오준성	오준성/은종태	nan
8	7	8위	연애소설	가비엔제이	민명기	민명기	nan
9	8	9위	Pretty Girl	카라	한재호/김승수	송수윤	nan
10	9	10위	좋은 날	백지영	방시혁	방시혁	nan
11	10	11위	붉은 노을	빅뱅	이영훈	이영훈/지드래곤	nan

Google docx Excel로 공동 수작업...

분류 모델 생성 및 성능

1. 데이터 전처리 과정에서 “tfidf vectorizer” 활용

가수, 작곡 각 단어의 빈도수를 Tfidf를 활용하여

가수/ 작곡 집합(사전)을 만들어서 벡터화

분류 모델 생성 및 성능

독립변수 : 가수, 작곡가, 음악장르

종속변수 : 순위 (1~100위의 연속적인 값을 예측)

2. "MLP" 활용

```
pred = mlp_clf.predict(x_data)
```

```
mean_squared_error(y_v_data, pred)
```

1123.7322814926122

3. "Linear Regression" 활용

```
from sklearn import datasets, linear_model  
model2 = linear_model.LinearRegression().fit(train_x, y_v_data)
```

```
pred = model2.predict(train_x)
```

```
from sklearn.metrics import mean_squared_error  
mean_squared_error(y_v_data, pred)
```

627.2924768855289

"Train set mean squared error" 값이 높음

에러값을 줄이는 것에 실패하여 목표를 재설정

분류 모델 생성 및 성능

TOP 10 가능성 예측

Top 10을 기준으로 순위를 두 카테고리로 나누어 True와 False값으로 반환

4. "Random forest" 활용

```
pred = model_rf.predict(train_x)
```

```
sum(df_2['순위_4'] == pred)/len(pred)
```

```
0.9014108022372486
```

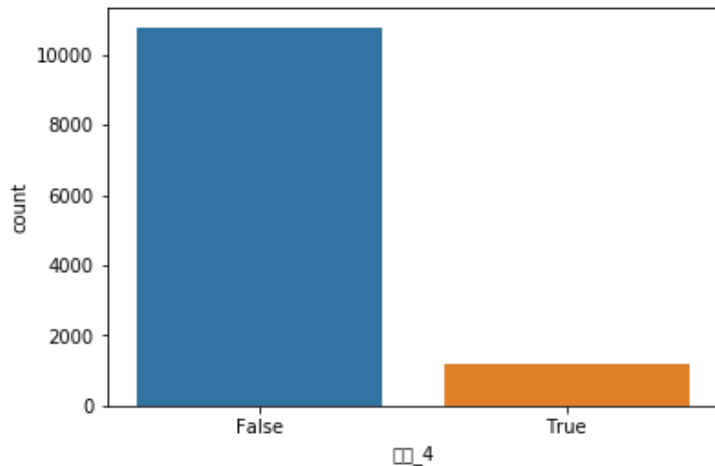
"Train set accuracy" 값이 90% 이상

의문점: 왜 이렇게 성능이 좋을까?

분류 모델 생성 및 성능

```
import seaborn as sns
sns.countplot(df_2['순위_4'])
```

<matplotlib.axes._subplots.AxesSubplot at 0xca84438>



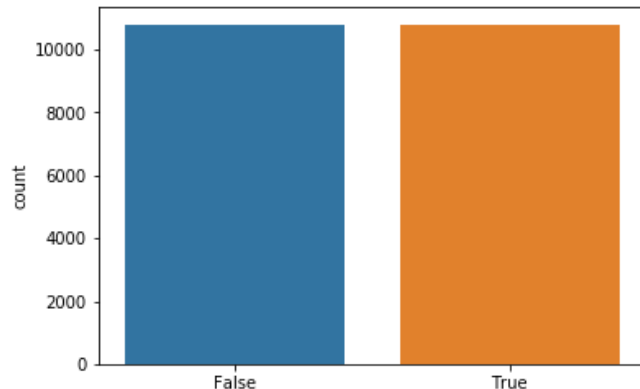
Label이 한 쪽으로 치우쳐져,
False만 답하도록 학습

```
#class 분포가 너무 치우쳐져 있다. => false라고만 대답하면 90프로는 맞는다..  
# 분포를 맞추기 위해 복사할 10위 안 데이터를  
import seaborn as sns  
train_top_10 = df_2[df_2['순위_2'] <= 10]
```

```
for i in range(8):  
    df_2 = pd.concat([df_2, train_top_10])
```

```
sns.countplot(df_2['순위_4'])
```

<matplotlib.axes._subplots.AxesSubplot at 0x14fe1a58>



1위부터 10위까지의 데이터를
8번 복사하여 데이터 생성

분류 모델 생성 및 성능

독립변수 : 가수, 작곡가, 음악장르

종속변수 : 순위 (TOP 10 True/False)

2. "Random forest" 활용

```
pred_test = model_rf.predict(test_x)
accuracy_score(test_y, pred_test)
```

0.6434

```
print(classification_report(test['순위_4'], pred_test))
```

	precision	recall	f1-score	support
False	0.62	0.73	0.67	2488
True	0.67	0.55	0.61	2512
avg / total	0.64	0.64	0.64	5000

3. "Logistic regression" 활용

```
from sklearn.linear_model import SGDClassifier
model_rf = LogisticRegression().fit(train_x, train_y)
```

```
pred_train = model_rf.predict(train_x)
accuracy_score(train['순위_4'], pred_train)
```

0.7350190091123047

```
from sklearn.metrics import classification_report
print(classification_report(train_y, pred_train))
```

	precision	recall	f1-score	support
False	0.79	0.64	0.71	8318
True	0.70	0.83	0.76	8253
avg / total	0.74	0.74	0.73	16571

Django 연동

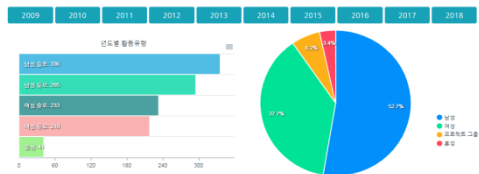


Process

- 1. 가수와 작곡가를 입력하세요.
- 2. 음악 장르를 선택하세요.
- 3. TOP 10 가능성 확인
아쉽네요! 10위안에 들 수 없을 것 같아요..
10위 안에 들 확률은 46% 입니다.
- 4. 10년 간 음악 산업 변화 데이터 시각화

Django 연동

활동유형 현황



활동유형 별 데이터 시각화

2009년 – 2018년 선택

Bar Chart : 여성 솔로 / 여성 그룹 /
남성 솔로 / 남성 그룹 / 혼성
Pie Chart : 여성 / 남성 / 프로젝트 그룹 / 혼성
프로젝트 그룹? 시즌 별 개인 가수 콜라보 그룹

년도별 재생시간 추이



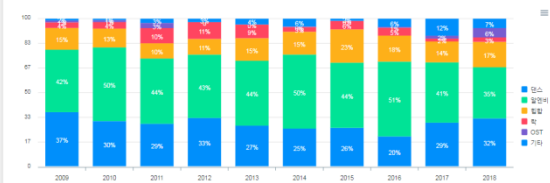
년도별 재생시간 추이

2009년 – 2018년

Candlestick Chart

: 남성 솔로 / 남성 그룹 / 혼성
여성 / 남성 / 프로젝트 그룹 / 혼성

년도별 음악장르 분포



년도별 음악장르 분포

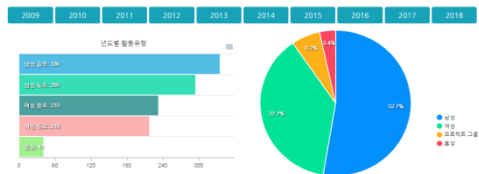
2009년 – 2018년

Stacked Column Chart

: 댄스 / 알앤비 / 힙합 / 락 / ost / 기타

Django 연동

활동유형 현황



활동유형 별 데이터 시각화

2009년 - 2018년 선택

Bar Chart : 여성 솔로 / 여성 그룹 /
남성 솔로 / 남성 그룹 / 혼성
Pie Chart : 여성 / 남성 / 프로젝트 그룹 / 혼성
프로젝트 그룹? 시즌 별 개인 가수 콜라보 그룹

년도별 재생시간 추이



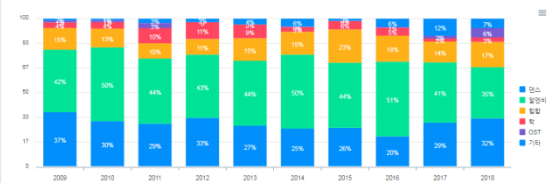
년도별 재생시간 추이

2009년 - 2018년

Candlestick Chart

: 남성 솔로 / 남성 그룹 / 혼성
여성 / 남성 / 프로젝트 그룹 / 혼성

년도별 음악장르 분포



년도별 음악장르 분포

2009년 - 2018년

Stacked Column Chart

: 댄스 / 알앤비 / 힙합 / 락 / ost / 기타

업무분장

팀장	유정오	프로젝트 전반 관리 및 분류 모델 생성
팀원	문경영	Python을 사용한 데이터 분석
팀원	최종찬	Python을 사용한 데이터 분석 및 데이터 전처리
팀원	황재훈	Django를 사용한 웹 코딩, Chart페이지 구현
팀원	홍다희	프로젝트 기획 및 데이터 전처리, 데이터 시각화
팀원	이대환	데이터 전처리

Q & A

Thank you