

닐로 날로잡기 프로젝트  
Database Project



Prof. 박재휘 교수님  
Project Manager 201401447 박재성  
201401421 남궁찬  
201401454 방기현  
201401500 최영훈

## 목 차(Table of Contents)

1. 개요
  - 1.1 프로젝트 주제
  - 1.2 프로젝트 개요 및 필요성
  - 1.3 프로젝트 목적
  - 1.4 기대효과
2. 프로젝트 일정
  - 2.1 프로젝트 일정표
3. 팀 구성 및 역할
  - 3.1 조직 구성
  - 3.2 역할
4. 사용 시나리오
  - 4.1 사용할 데이터
  - 4.2. 사용할 데이터 확보 방법
5. 개발 시나리오
  - 5.1 개발 시나리오
  - 5.2 조작 음원 판단 기준
  - 5.2 흐름도
  - 5.3 데이터베이스 모델(ERD)
6. 개발 환경
  - 6.1. 데이터베이스
  - 6.2 데이터 분석
  - 6.3 머신러닝
  - 6.4 데이터 수집
  - 6.5 협업, 버전관리
7. 개발 과정에서의 유의점
  - 7.1. 알고리즘 구현 난이도
  - 7.2. 데이터 확보의 어려움
8. 최근 동향
  - 8.1. 최근 동향
  - 8.2. 관련 기사
  - 8.3. 관련 연구
9. 참고문헌

## 1. 개요

### 1.1 주제 - 음악 플레이어 데이터 분석을 통한 순위 조작여부 예측

### 1.2 프로젝트 개요 및 필요성

음원 사이트들은 음원 발매 이후 한 시간 단위로 집계되는 ‘실시간 차트’를 발표하고 있다. 최근 음원 문화의 변화에 따라 음원차트의 상위권에 노출되지 못하면 사장되어버린다는 위기의식이 콘텐츠 제작자와 극성팬들 사이에 팽배하면서 일부 이해관계자들의 불안 심리를 이용해 음원차트의 공정성에 악의적인 위해를 가하고자 행해지는 음원 사재기가 문제가 되고 있다. 실제로 2015년 9월 말 국내 최대 음원 사이트인 멜론이 각 음반 기획사를 대상으로 ‘[로엔] 음원사재기 (어뷰징) 행위 근절을 위한 협조 요청의 건’이라는 제목의 공문을 보낸 사실이 드러났으며, 공문에는 음원 사재기를 이용한 순위 차트 조작에 유감을 표하며 외부 업체와의 부당한 사례가 있을 경우 법률적 조치를 취하겠다는 내용이 담겨 있었다. 이에 대응하여 문화체육관광부 주관으로 2016년 3월 3일 ‘음악 산업 진흥에 관한 법률’을 통과시켰다.

위와 같은 이유로 음원 차트 데이터를 분석하여 사재기로 의심되는 음원을 알아내는 프로젝트를 수행하게 되었다.

### 1.3 프로젝트의 목적

프로젝트의 목적은 지금까지의 음원차트와 트렌드 분석을 통해 조작으로 의심되는 음원들을 찾아내고 해당 음원의 패턴을 알아내는 것이다. 나아가 알아낸 패턴을 통해 새로운 음원에 대한 조작 의심 정도를 알려주는 머신러닝 모델을 만드는 것이다.

### 1.4 기대효과

머신러닝 모델을 통해 나온 조작 의심 정도를 음원 차트 순위 알고리즘에 패널티로 적용한다면 좀 더 공정하고 신뢰할 만한 음원 차트를 제공할 수 있을 것이다. 이는 음악 산업의 교란을 근절하고 음악 산업 진흥을 이끌어 낼 수 있을 것이다.

	9월					10월					11월					12월				
	3	10	17	24	31	1	8	15	22	29	5	12	19	26	30	3	10	17	24	
PROJECT WEEK	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
기획서 초안																				
	조직 구성																			
		주제 선정					피드백													
			조안 작성																	
기획서 최종안								피드백 회의												
									최종안 작성											
									발표											
프로젝트 구현							RDS시작													
								DB설계						DB보안 수정						
									구현 및 회의											
									알고리즘 설계				알고리즘 보안 수정							
														인터페이스 구현						
보고서 초안												초안 작성								
												피드백 회의								
보고서 최종안															최종안 작성					
																발표 및 시연				
프로젝트 완료																				

### 3. 팀 구성 및 역할

#### 3.1 조직 구성

한 학기 동안의 프로젝트 인 것과 프로젝트의 규모를 고려하여, 빠른 의사소통과 결정을 위해서 분산형 팀조직으로 조직을 구성하였다.

- 1)민주주의식 의사결정
- 2)같이 협동, 수행하는 비이기적인 팀
- 3)자신의 역할에 맞는 일을 알아서 수행

#### 3.2 역할

구성원		할당된 작업 및 역할
조원명	*박 재 성	Project Manager, 전반적인 프로젝트 관리, 각종 지원, 자료수집
	남 궁 찬	Question & Answer, 자료수집, 기술 파악 및 수집, 동향분석
	방 기 현	Development Manager, DBA, 코드 기획 및 개발 ,자료 수집
	최 영 훈	Document Manager, 기획-보고서 초안 및 최종안 제작, 발표자료 제작

### 4. 사용 데이터

#### 4.1. 사용할 데이터

지니, 엠넷, 벅스 3곳의

- 시간대별 음원순위 : 각 시간대별 음원순위 데이터를 가져와 변동폭 등을 분석한다.
- 해당 음원의 화제성 : 음원순위와 해당 음원의 화제성 관계를 분석한다.
- 아티스트의 유명세 정도 : 음원순위와 해당 음원 아티스트의 유명세 정도 관계를 분석한다.

총 5가지의 데이터를 사용하려 합니다.

#### 4.2. 사용할 데이터 확보 방법

- 시간대별 음원순위 : 지니, 엠넷, 벅스 세 곳의 시간대별 음원차트를 크롤링하여 확보한다.
- 해당 음원의 화제성 : 구글 트렌드 오픈 API를 이용한다.
- 아티스트의 유명세 정도 : 1)구글 트렌드 오픈 API를 이용한다.

---

1) <https://trends.google.co.kr/trends/>  
<https://github.com/GeneralMills/pytrends>

## 5. 개발 시나리오

### 5.1. 개발 시나리오

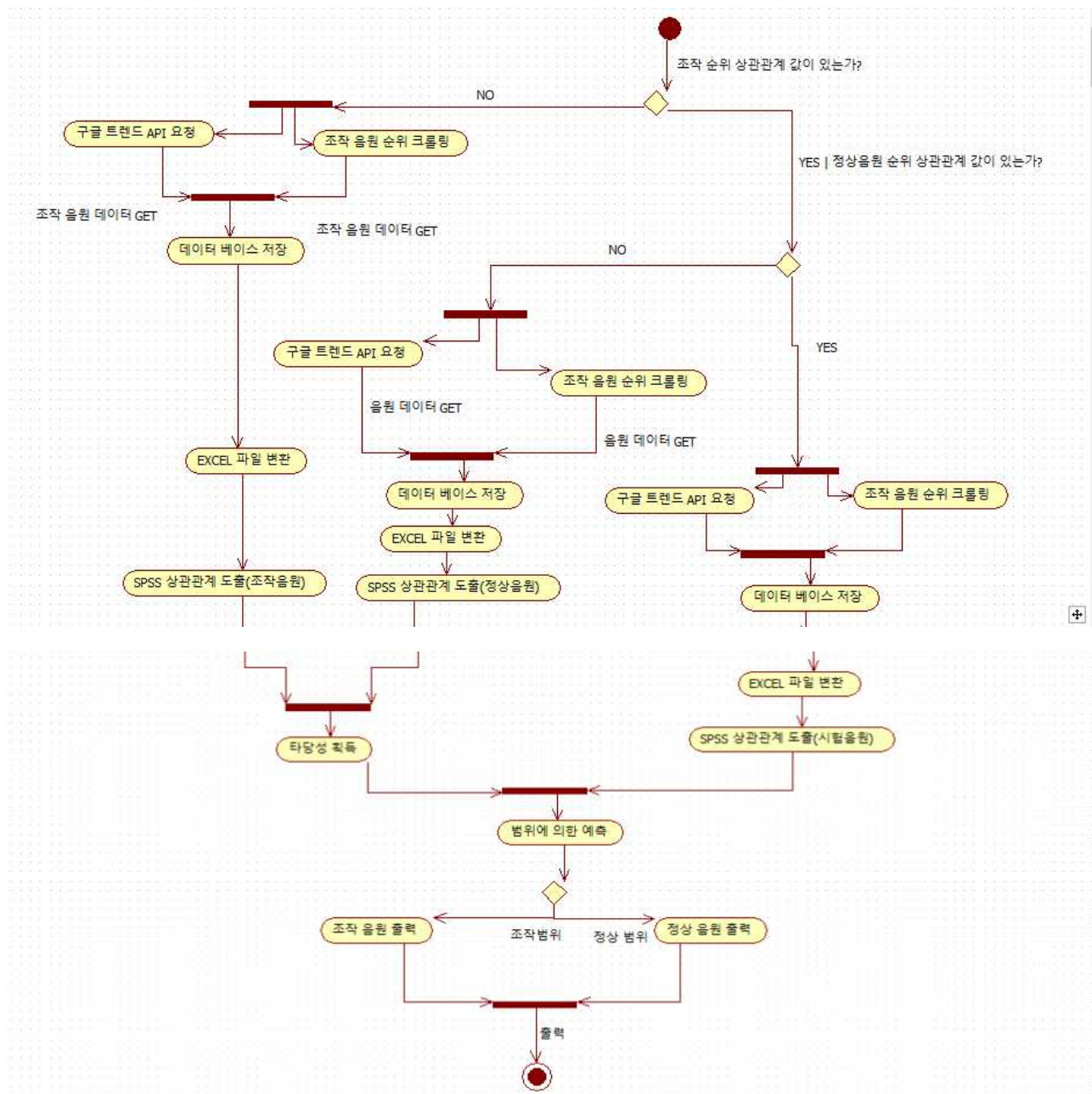
- 1) 벅스, 지니, 엠넷의 음원차트를 파싱하여 데이터베이스에 저장한다.
- 2) 구글 API를 사용하여 음원에 해당하는 키워드의 트렌드 지수를 데이터베이스에 저장한다.
- 3) 데이터베이스에서 데이터를 가져와 분석한 뒤 조작으로 의심되는 음원을 찾아낸다.
- 4) 해당 음원의 특정 패턴을 찾아낸다.
- 5) 찾아낸 패턴을 이용하여 조작 의심 정도를 산출하는 머신러닝 모델을 만든다.
- 6) 테스트 & 개선

-높은 신뢰성을 시스템이 나올 때 까지 반복

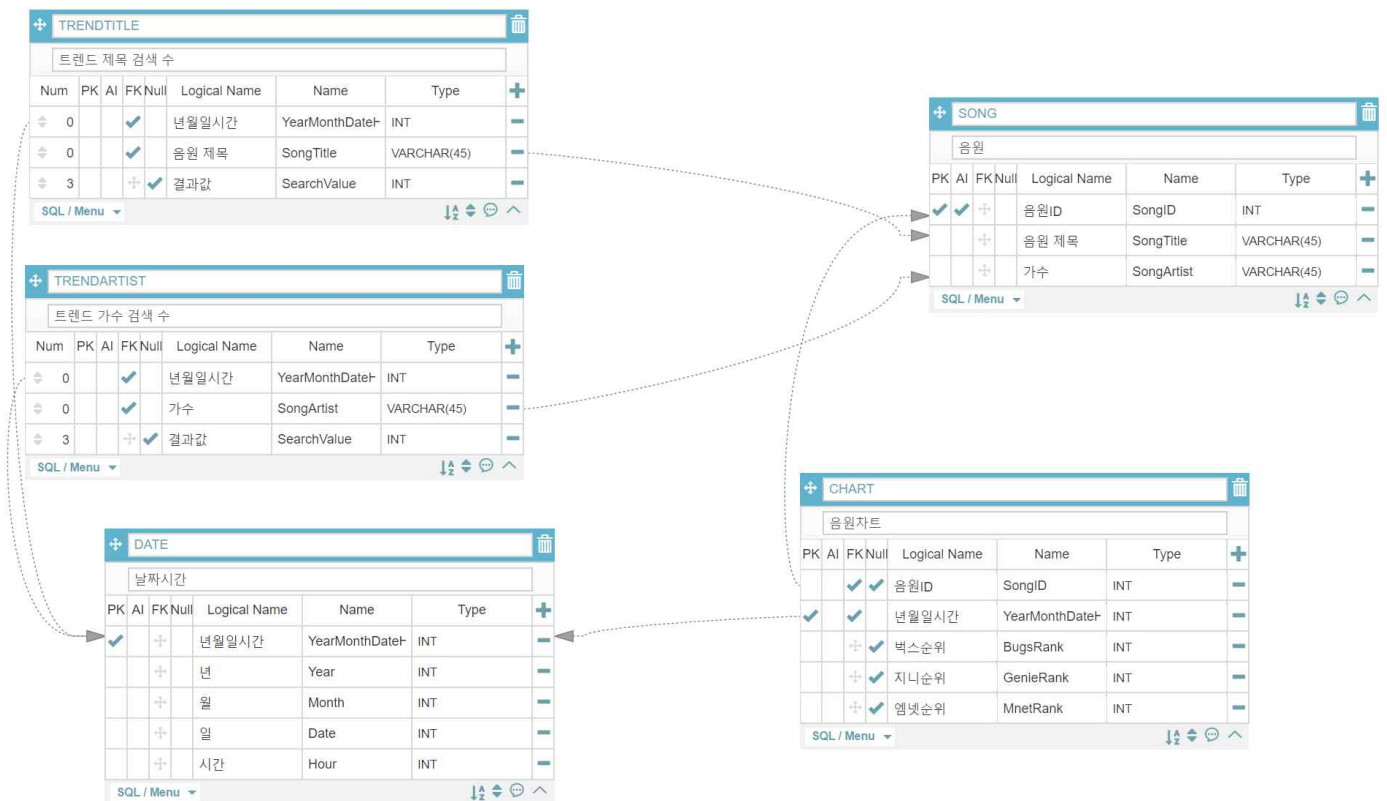
### 5.2 조작 음원 판단 기준

정상적인 음원은 음원 차트의 순위의 움직임과 구글 검색어 트렌드의 검색 빈도의 움직임이 높은 상관관계를 가질 것이다 라는 가정(검색 빈도와 순위의 움직임은 대체적 상관관계가 높음을 이용)을 시작으로, 조작 음원 차트는 차트 순위를 올리기 위한 가짜 계정 스트리밍, 다운로드 횟수의 증가로 비정상적인 순위 움직임이기 때문에, 구글 검색어 트렌드의 검색 빈도와는 낮은 상관관계를 가질 것이다. 라는 두 번째 가정을 통계 프로그램을 이용한 타당성 획득으로 조작 음원의 기준 잡을 것입니다.

## 5.3 흐름도



## 5.4. 데이터베이스 모델(ERD)



## 6. 개발 환경

### 6.1. 데이터베이스

AWS의 RDS중 Mysql을 사용한다.

### 6.2 머신러닝

Scikit-Learn 기술을 이용하여 머신러닝 모델 생성

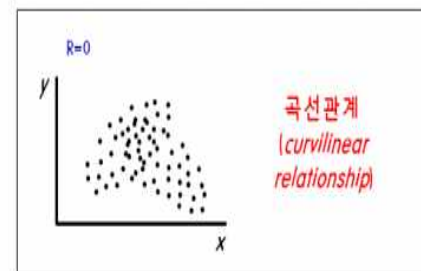
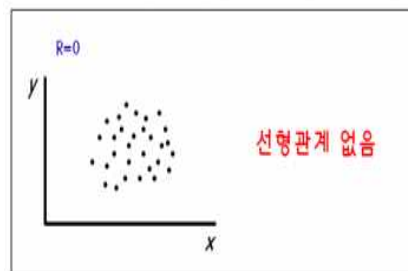
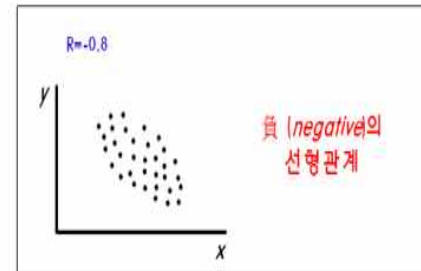
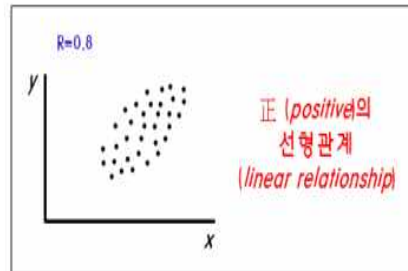


## 6.3 데이터 분석

IBM SPSS Statistics 와 python을 이용하여 데이터 분석

데이터를 다음과 같이 csv확장자파일에 정렬한 다음, IBM SPSS를 통해 상관관계를 도출  
SPSS에서 R이 1에 가까우면 다음과 같이 긍정선형관계 혹은 부정선형관계 가 형성, 즉 X와 Y  
가 서로 연관이 있다는 뜻이며, 데이터들의 타당성을 획득 할 것입니다.

1	카테고리: 모든 카테고리		
2			
3	일	트렌드	박스
4	2017-11-28	30	0
5	2017-11-29	31	0
6	2017-11-30	62	27
7	2017-12-01	64	55
8	2017-12-02	34	40
9	2017-12-03	0	25
10	2017-12-04	0	46
11	2017-12-05	31	23
12	2017-12-06	30	38
13	2017-12-07	0	32
14	2017-12-08	32	22
15	2017-12-09	34	20
16	2017-12-10	0	38
17	2017-12-11	31	37
18	2017-12-12	0	50
19	2017-12-13	0	42
20	2017-12-14	0	40
21	2017-12-15	0	54
22	2017-12-16	0	41
23	2017-12-17	33	43
24	2017-12-18	0	53
25	2017-12-19	31	55



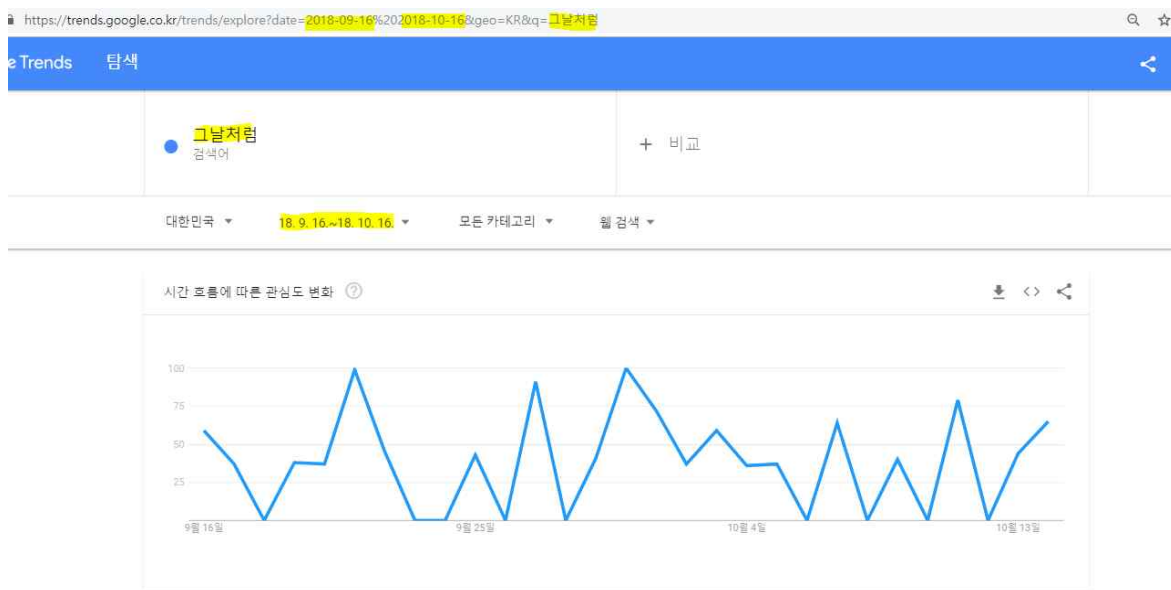
## 6.4 데이터 수집

### 1) 자바와 파이썬을 이용한 데이터 크롤링을 이용한 파싱(데이터 수집 자동화)

The screenshot shows the music.bugs.co.kr website with a real-time chart for the date 2018-10-16 at 22:00. The chart lists songs, with '몇지게 인사하는 법 (Feat. 슬기 of Red Velvet)' by Zion.T at the top. The right sidebar shows the HTML structure of the chart, including a table with song details.

순위	곡	아티스트	앨범	듣기
1	몇지게 인사하는 법 (Feat. 슬기 of Red Velvet)	Zion.T	ZZZ	

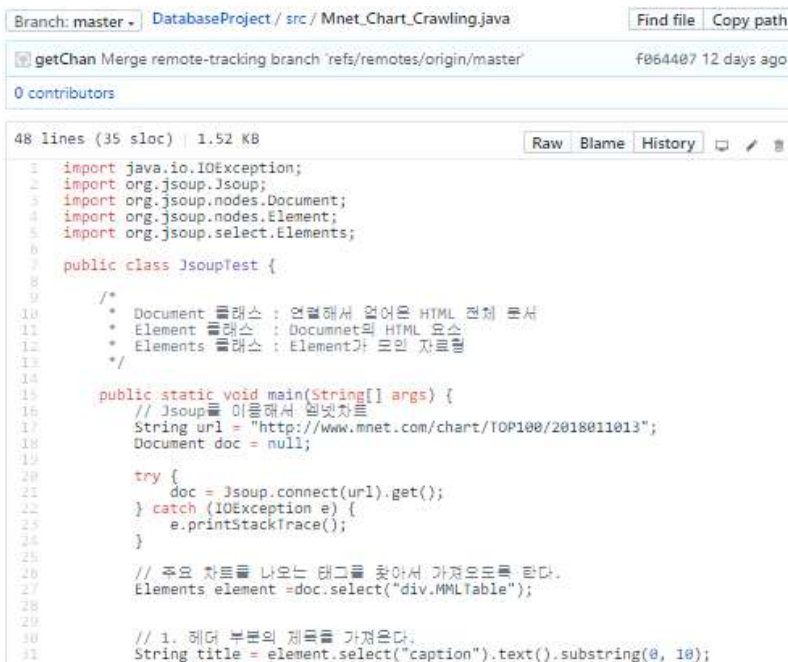
### 2) 자바와 파이썬을 이용한 구글 트렌드 오픈 API 사용



## 6.5 협업, 버전관리

### Git, Github 사용

#### 1) 정형화 된 코드 공유



```
Branch: master DatabaseProject / src / Mnet_Chart_Crawling.java Find file Copy path
getChan Merge remote-tracking branch 'refs/remotes/origin/master' f064407 12 days ago
0 contributors

48 lines (35 sloc) 1.52 KB Raw Blame History

1 import java.io.IOException;
2 import org.jsoup.Jsoup;
3 import org.jsoup.nodes.Document;
4 import org.jsoup.nodes.Element;
5 import org.jsoup.select.Elements;
6
7 public class JsoupTest {
8
9     /*
10     * Document 클래스 : 연결해서 얻어온 HTML 전체 문서
11     * Element 클래스 : Document의 HTML 요소
12     * Elements 클래스 : Element가 모인 자료형
13     */
14
15     public static void main(String[] args) {
16         // Jsoup을 이용해서 웹페이지
17         String url = "http://www.mnet.com/chart/TOP100/2018011013";
18         Document doc = null;
19
20         try {
21             doc = Jsoup.connect(url).get();
22         } catch (IOException e) {
23             e.printStackTrace();
24         }
25
26         // 주요 차트를 나오는 태그를 찾아서 가져오도록 한다.
27         Elements element = doc.select("div.MVTable");
28
29         // 1. 헤더 부분의 제목을 가져온다.
30         String title = element.select("caption").text().substring(0, 10);
31     }
32 }
```

#### 2) 회의 내용 정리



### 3)참고 자료 모음

🔍	🔍 3 Open ✓ 0 Closed	Author +	Labels +	Projects +	Milestones +	Assignee +	Sort +
🔍	🔍 프로젝트 기획서 모범안	#3 opened 7 days ago by dudgns3tp					
🔍	🔍 Python을 이용한 구글 트렌드 크롤링	#2 opened 12 days ago by wotjd4305					
🔍	🔍 Jsoup를 이용한 크롤링 법	#1 opened 12 days ago by wotjd4305					

🔍 ProTip! Exclude your own issues with `-author:wotjd4305`.

## 7. 개발 과정에서의 유의점

### 7.1. 알고리즘 구현 난이도

데이터간의 상관관계와 데이터의 특정한 패턴을 파악하는 것이 핵심이며, 제대로 구해졌을 경우, 알고리즘 구현에 있어 큰 어려움이 없을 것으로 예상됨.

머신러닝 모델 구현의 경우 조작으로 의심되었던 음원의 수가 많지 않아 정교한 모델링이 필요할 것으로 보여짐.

또한 예외상황(버스커버스커 - 벚꽃엔딩과 같은 계절 노래, 혹은 재조명) 발생에 대한 알고리즘의 세부적인 구현이 필요하다.

음원 조작 판단의 기준을 삼기 위해 IBM SPSS로 타당성을 위한 상관관계를 얻고자 하였을 때, 얻고자하는 신뢰성이 나오지 않았을 때가 발생 가능 할 수 가 있다. 따라서 높은 신뢰성(타당성)을 위한 반복적인 알고리즘 재구현의 과정이 필요할 것으로 예상됨.

마지막으로 만약 이 알고리즘을 바탕으로 프로젝트가 끝나고, 실제로 상용화가 되었을 때, 프로젝트의 알고리즘을 파악하고 음원 조작을 하는 브로커들이 구글 트렌드의 검색빈도를 의식하여 검색빈도를 같이 올리는 작업을 수행 시, 이를 해결하기 위한 개선이 필요할 것으로 예상되며, 알고리즘의 복잡도가 더욱 증가할 것으로 예상됨.

### 7.2. 데이터 확보의 어려움

음원사이트들이 공개적인 API를 제공하고 있지 않다. 따라서 크롤링을 이용하여 원하는 데이터를 가지고 오는 작업이 필요하다. 크롤링 알고리즘을 각 음원순위 웹페이지에 맞게 구현하여 가져오는 작업에 어려움은 없을 것이라 보여진다.. 이 외의 데이터는 구글에서 제공하는 API혹은 오픈소스 API를 이용하여 데이터를 확보할 수 있다. 가져온 데이터들을 알맞게 수치화하는 등 정리하는 과정에서 어려움이 예상된다.

## 8. 최근 동향

### 8.1. 최근 동향

2018년 4월 리메즈엔터테인먼트 소속 가수 닐로가 비정상적인 음원 그래프를 보이며 음원 사재기 의혹에 휩싸였다. 같은 소속사 가수들 또한 의혹에서 벗어날 수 없게 되었다.

2018년 7월에는 손의 way back home이 음원사재기 의혹을 받고 있으며 이와 함께 리메즈의 소속 가수들의 사재기 논란이 재조명받게 되었다. 이에 네티즌들은 '손 안대고 닐로 먹기'라는 신조어를 탄생시켰다.

2018년 8월 말에는 오반의 '20살이 왜이렇게 능글맞아'가 7시간 만에 10위권으로 올라와 논란이 되고 있다.

2018년 10월 10일에는, 아무런 계기없이 리메즈엔터테인먼트의 반하나가 자이언티X슬기(레드벨벳), 양다일, 아이유, 임창정, 로꼬, 신곡이 계속 나오는 쇼미더머니 7 등의 역대급 음원강자들이 포진해있는, 그야말로 과포화상태인 멜론차트에서 7,8시 미진입, 9시 91위, 10시 74위, 11시 32위 12시 24위를 하는등의 이상한 추이를 보이고 있다. 타 차트에서도 급격한 상승을 보이고 있다. 논란이 확실시 된 결정적 이유는 음원 발매 후 2시간동안 진입을 하지 않았다가, 다른 음원들의 순위가 크게 낙폭하고 있음에도 거의 수직상승에 가깝게 순위를 높였기 때문이다. 대중성이 확립된 가수가 음원 발매를 할 경우엔, 초반 차트 진입을 했다가 서서히 떨어지는게 정상적인 추이이다. 하지만 이 가수의 곡은 차트 프리징이 시작될 새벽 1시 타임까지 순위가 비정상적으로 급상승하기 시작하고, 이건 매우 비정상적인 추이가 될 수 밖에 없다.

### 8.2. 관련 기사

김 의원이 문체부로부터 제출받은 자료에 따르면, 문체부는 업계 관계자 신고 등에 따라 가수 '닐로'와 '손'의 음원을 둘러싼 사재기·차트 조작 의혹 사건을 4월과 6월 각각 차례로 접수했다. 문체부는 사건 접수 후 지니뮤직·멜론·벅스뮤직·네이버뮤직·엠넷·소리바다 등 6대 음원서비스 사업자들에게 관련 자료 제출을 요구했지만, 10월1일까지 지니뮤직의 자료만 확보한 것으로 확인됐다고 김 의원 측은 설명했다.

나머지 5개 사업자는 이달 중순 자료 제출을 하겠다고 문체부에 통보한 것으로 전해졌다. 문체부는 이와 별도로 지난 8월 음원 사재기 관련 데이터 분석 용역을 발주했지만, 올해 12월 말에야 결과를 받아보기로 하는 등 '거북이 대응'을 하고 있다고 김 의원은 지적했다.(10/8)

한편 문체부는 지난 4월 리메즈의 요청으로 시작한 닐로 관련 사재기 의혹 건과 이후 추가된 손건에 대한 조사 결과를 다음달 중 발표할 예정이다. 문체부 관계자는 “6개 음원 업체 중 4곳으로부터 전체적인 자료를 넘겨받았고 1곳은 일부 자료를 제출했다”며 “제출받은 데이터에 대해서는 분석 작업을 진행하고 있다”고 말했다.(10/16)

### 8.3. 가온차트 연구 결과



gaon chart

닐로의 역주행과 기존 역주행의 차이점은...

닐로의 '지나오다'는 첫째, 별다른 이슈 없이 역대 최단 시간에 1위에 오른 역주행 곡이 될 것으로 예상됩니다. 지난 2014년 브로의 '그런 남자'의 경우 출시 후 주간 차트 기준 38위에서 곧장 1위에 오른 사례가 있긴 합니다. 그러나 브로의 '그런 남자'는 '약을 먹었니', '연봉 육천' 등 직설적이고 현실적인 가사로 남성들에게는 공감과 여성들에게서는 반감을 사며 논란이 확산되어 단숨에 1위에 올랐던 케이스입니다.

둘째, 기존 역주행곡들에서 나타나는 부침의 과정, 즉 바닥을 다지면서 순위가 상승하는 모습이 관찰되지 않습니다. 음원 역주행시 보통 나타나는 일시적 순위 하락 또는 횡보 후 재상승 등의 과정을 닐로의 곡에서는 찾아보기 어렵습니다.

마지막으로, 기존 역주행 곡들에서 나타나는 역주행을 유발할 만한 직접적인 사건과 계기를 찾기 어렵습니다. 앞서 살펴본 EXID는 '하니 직캠', 한동근은 '커버 동영상', '라디오 스타 출연', '듀엣가요제 출연', 윤종신은 '세로 라이브', '유희열의 스케치북 출연', 걸그룹 여자친구는 '파당 사건' 등 역주행의 원인이 되는 구체적인 사건들이 존재했었습니다.

김진우 가온차트 수석연구위원

## 9. 참고문헌

곽아영,(2017) 데이터 분석을 통한 온라인 음원차트의 어뷰징 영향요인 탐색, *학위논문(석사) 서  
울시립대학교 대학원 : 경영학과 2017. 2*

가온차트, ‘닐로 사태’ 팩트 체크, 김진우 수석연구위원

<http://www.gaonchart.co.kr/main/section/article/p.view.gaon?idx=13879>

나무위키, 음원 사재기

<https://namu.wiki/w/%EC%9D%8C%EC%9B%90%20%EC%82%AC%EC%9E%AC%EA%B8%B0>

### 관련기사

<http://www.nocutnews.co.kr/news/5039567>

[http://www.seoul.co.kr/news/newsView.php?id=20181016500051&wlog\\_tag3=naver#csidxb6b468085370be489d6a1ccc2053913](http://www.seoul.co.kr/news/newsView.php?id=20181016500051&wlog_tag3=naver#csidxb6b468085370be489d6a1ccc2053913)