닐로 날로잡기 프로젝트

Database Project

Prof.박재휘

Project Manager 201401447박재성

201401421남궁찬

201401454방기현

201401500최영훈

1.주제

– 음악 플레이어 데이터 분석을 통한 순위 조작 여부 예측

2.개발동기(시나리오의 영향력/필요성)

– 음원 중 인지도가 낮은 가수의 노래가 뒤늦게 화제가 돼 상위권에 오르는 역주행이 종종 있습니다. 그런데 최근 들어 예전 역주행과는 다른 전례 없는 역주행 사례가 발생하고 있습니다. 이를 음원차트 조작이라고 하는데 이러한 현상이 조작이 아니라면 문제가 되지 않겠지만 만약 조작이라면 이것은 가요계의 공정성과 다양성을 해칠 수 있는 큰 문제가 될 수 있습니다. 이러한 문제를 해결하기 위해 우리는 음원 사이트들의 데이터를 분석을 통하여 이를 찾아낼 방법이 있을 거라 판단하였고 이 프로젝트를 진행하기로 하였습니다.

[참고 자료]음원 사재기의 도출방법 http://impossibleproject.tistory.com/4169

3.개발 진행 방향.

3-1-1) 사용할 데이터의 상세 설명

지니, 엠넷, 벅스 3곳의

-시간대별 음원순위 : 각 시간대별 음원순위 데이터를 가져와 변동폭 등을 분석합니다.

-해당 음원의 화제성 : 음원순위와 해당 음원의 화제성 관계를 분석합니다.

-아티스트의 유명세 정도 : 음원순위와 해당 음원 아티스트의 유명세 정도 관계를 분석합니다.

-소셜미디어 조회수 : 유튜브, 트위터, YinYueTai의 소셜미디어 조회수와 음원순위와의 관계를 분석합니다.

-연령대별 음원순위 : 연령대와 음원순위와의 관계를 분석합니다.

총 5가지의 데이터를 사용하려 합니다.

3-1-2)해당 데이터를 확보하는 방법이나 전략

-시간대별 음원순위 : 지니, 엠넷, 벅스 세 곳의 시간대별 음원차트를 파싱하여 확보합니다.

-해당 음원의 화제성 : 네이버 통합검색어 트렌드 API를 이용합니다.

-아티스트의 유명세 정도 : 네이버 통합검색어 트렌드 API를 이용합니다.

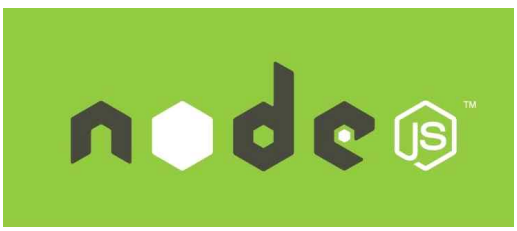
-소셜미디어 조회수 : 가온 소셜차트(유튜브, 트위터, YinYueTai의 데이터를 수집하여 집계)를 파싱하여 확보합니다.

-(+연령대별 음원순위는 확보 방법을 찾아보는 중입니다.)

-네이버트렌드(https://datalab.naver.com/keyword/trendSearch.naver)

3-2) 개발 도구, 방향:

-서버



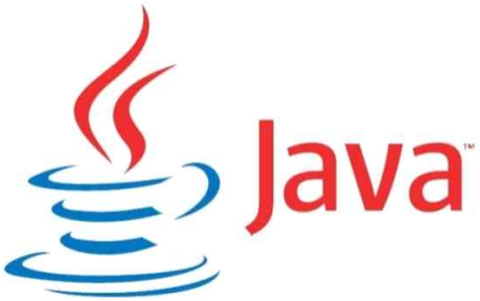
아마존 웹 서비스를 이용하여 node.js로 서버 구축, 서버에서 음원사이트 데이터를 파싱하여 Datebase에 저장.

– 데이터베이스

mysql 관계형 데이터베이스 이용.

- **머신러닝 기술**
  데이터베이스로부터 머신러닝 모델을 생성하고
  음원조작이 의심되는 객체 추출.

- **웹 크롤링**
자바와 파이썬을 통하여 데이터를 크롤링.



- **데이터 분석**
IBM SPSS Statistics 와 python을 이용하여 데이터 분석

3-3) 알고리즘 계획(개발 시나리오)

- 우리의 계획은 음악 플레이어의 순위를 매기자는 것이 아니며, 플레이어의 급격한 순위 변동이 정상적인 움직임인가를 알아보고자 하는 것이다. 따라서 우리는 데이터 분석 프로그램을 이용하여 조사한 데이터중 어떤 데이터가 음원 차트 순위에 가장 큰 영향을 주면서, 그 데이터 요소가 순위 결정에 상관이 있는지 확인하고, 도출된 데이터의 상관관계 요소와 신뢰도를 기반으로 조작 여부를 예측하는 것이 핵심이다.

3-4) 구현 방법

1)대부분 음원사이트들은 API를 제공하지 않기 때문에, 사이트 외부에서 제공하는 정보들을 구축한 데이터베이스에 파싱하여 저장한 것이 첫 단계
2)음원순위 이외의 속성들(화제성 등)을 검색엔진 API를 이용해 가져온 뒤 적절히 수치화하여 데이터베이스에 저장
3)데이터베이스에 저장해둔 데이터를 분석하여 음원의 순위를 결정하는 데이터들의 관계를 분석 파악한다.
4) 조작이 의심되는 음원(닐x, 숀x, 장x철)들의 데이터의 특징을 파악,
5) 머신러닝 모델을 구현.
6) 데이터를 입력하여 원하는 결과가 나오는지 확인검토 및 수정.

3-5) 개발 과정에서의 유의점(프로젝트 수행에서 예상되는 어려움)

1) 알고리즘 구현 난이도 - 데이터들의 통계로 근거 있는 상관관계를 통해서 알고리즘을 구현하여야 합니다.
따라서 저희는 데이터간의 상관관계를 파악하는 것이 핵심이며, 상관관계가 제대로 구해졌을 경우, 알고리즘 구현에 있어 큰 어려움이 없을 것으로 예상됩니다.
2) 데이터 확보의 어려움 - 여러 음원사이트들이 공개적인 API를 제공하고 있지 않습니다. 그래서 저희는 위에서 언급한 크롤링을 이용하여 저희가 원하는 데이터를 가지고 오는 작업이 필요합니다. 크롤링 알고리즘을 각 음원순위

웹페이지에 맞게 구현하여 가져오는 작업에 어려움은 없을 것이라 판단이 됩니다. 이 외의 데이터는 네이버 혹은 구글에서 제공하는 API를 이용하여 데이터들을 수집할 수 있습니다. 가져온 데이터들을 수치화하는 등 정리하는 과정에서 약간의 어려움이 예상됩니다.

3) 음원조작 여부 판별를 최초로 시도 - 최근 음원조작이 많은 이슈를 낳고 있습니다. 음원차트의 급격한 변화와 화제성이 없음에도 불구하고 차트에 진입하는 등 조작을 의심할만한 근거와 자료들은 많이 찾아볼 수 있지만, 많은 데이터를 분석하면서 체계적으로 접근하는 방식은 찾아볼 수 없었습니다. 따라서 본 프로젝트 팀은 의심할 만한 여러 가지 요소를 추출하고 음원차트와의 상관관계를 분석하여 왜 해당 음원이 조작되었는지의 근거를 데이터 분석으로 찾아내는 것이고, 나아가서 음원조작 여부를 판별할 수 있는 머신러닝 모델을 구축하는 것이 목표입니다.

4) 프로젝트 개발의 이전 과 이후 - 음원사이트의 순위는 단순한 순위를 넘어선 문화이며, 음원사이트에서 제공받는 순위는 사람들이 받는 엄연한 서비스입니다. 따라서 이용자들이 받는 서비스는 정확해야합니다. 하지만 순위 조작으로 인해 이용자들은 부정확한 서비스를 받게 되었고, 순위라는 정보는 더 이상 정확성과 유용성을 갖지 못하게 되었습니다. 저희 프로젝트 개발이 성공적으로 마무리가 된 이후에는 순위라는 문화, 서비스를 의심 없이 이용할 수 있는 수준 이상으로 만드는 것에 유의할 것입니다.