

[인공지능과 기계학습 기말 대체 프로젝트 최종 보고서]

[프로젝트 주제 : 음원차트 실시간 스트리밍 정보를 활용한 음원 분류]



보고서 제출 날짜 :

조 이름 : 이하진

조 구성원 : 이하진

**요 약**

|  |
| --- |
| (과제의 내용, 결론(분석 결과 및 시사점)을 1페이지 이내로 작성)  본 프로젝트는 스트리밍 정보를 활용해 음원차트에 있는 다양한 음원들을 “대중형 음원”, “팬덤형 음원”, “사재기 의심 음원”으로 분류하는 것을 목표로 한다.  “가이섬.com”에서 음원 정보를 직접 크롤링하였으며, 데이터 정제를 한 후 음원 분류를 위해 비지도 학습방법인 K-Means Clustering과 Agglomerative Clustering을 활용하였다.  분석 결과, 음원 실시간 스트리밍 양의 변동성을 활용하여 데이터를 “대중형 음원”과 “팬덤형 음원”으로 분류할 수 있었으나 사재기 음원”을 분류하는 것에는 실패하였다. 또한, 분석 결과로 현재 음원시장에 “팬덤형 음원”의 비율이 과도하게 높다는 결론을 도출할 수 있었다.  각 단계에 대한 자세한 설명은 보고서 본문에서 찾아볼 수 있다. |

**목 차 (보고서 페이지에 맞게 아래 수정)**

|  |
| --- |
| 1. 서론 ……………………………………………………………………………………………………………………… 3  2. 기존 사례 …………………………………………………………………………………………………………… 4  3. 과제 수행 방법, 절차 및 내용 …………………………………………………………………………5  4. 과제 수행 결과 분석 …………………………………………………………………………………………6  5. 결론(분석 결과 및 시사점)…………………………………………………………………………………7  6. 참고자료 ………………………………………………………………………………………………………………8  7. 팀원명 및 역할 ……………………………………………………………………………………………………9 |

1. **서 론**

|  |
| --- |
| (주제 선정 이유, 수행 목적, 필요성을 논리적으로 설득력 있게 작성)  2018년쯤부터, 몇몇 가수들의 음원이 비정상적인 추이를 보이며 급격하게 차트 순위가 상승하는 현상이 반복되면서 음원 사재기 논란이 화두에 올랐다. 또한, 예전부터 문제로 지적되어 왔던 대형 팬덤의 “스트리밍 공세”(팬들이 여러 기기를 사용해 스트리밍을 24시간 동안 켜 두면서 자기 가수의 음원성적을 올리는 것)도 여전했다. 이러한 상황 속에서, 더 이상 음원차트가 대중의 선호도를 반영하지 못한다는 비판이 나왔다. 그래서 실시간 사용자 추이를 분석하여 음원차트에 있는 여러 음원들을 “대중형 음원”, “팬덤형 음원”, “사재기 의심 음원” 등으로 분류하고, 이를 통해 현재 음원 차트가 얼마나 대중의 선호도를 반영하는지 알아보고자 한다. |

1. **기존 사례**

|  |
| --- |
| (기존에 유사한 프로젝트나 연구가 무엇이 있는지 작성하고 본 과제의 차별성 제시)  조사 결과, 해당 주제에 대한 프로젝트나 연구가 거의 없었다. 음원 차트 관련 프로젝트는 대부분 음원 순위 혹은 음원 흥행 정도를 예측하는 것이었고, 관련 연구는 음원 사재기 현상을 막기 위해 정부가 음악산업에 개입해야 하는지에 대한 것이었다. 실제 데이터를 분석해 음원 종류를 분류하는 것은 아직 시도가 많이 되지 않은 듯 하다. |

1. **과제 수행 방법, 절차 및 내용**

|  |
| --- |
| (데이터 정제, 분석 기법, 해당 분석 기법 선택한 이유, 분석 과정을 반드시 포함)  \*중간대체과제 제출 당시와 비교하여 주제가 변경되었거나 수집 데이터가 바뀌었을 경우 사유와 데이터 수집 과정도 작성 필요  본 프로젝트를 위해 국내 음원 차트 아카이빙 사이트인 “가이섬.com”에서 Genie 탑100 음원차트 정보를 크롤링했다. 국내 최대 규모 음원 사이트인 Melon의 데이터를 사용 싶었으나, 새벽 1시부터 아침 7시까지의 데이터가 제공이 안되는 관계로 국내 두번째 규모의 음원사이트인 Genie의 음원 정보를 활용했다. 크롤링할 당시 데이터를 파이썬의 default dictionary 형태로 저장하였다. Key는 “곡 제목\_가수 이름”으로 설정하였고, 해당 곡의 데이터를 {날짜&시간: 값, 누적 청취자: 값, 누적 스트리밍 수: 값}의 형태로 모아서 list에 저장해 value로 설정했다. 모은 데이터는 로컬에 json 파일로 저장했다.  데이터를 정제하기 위해 우선 이상 데이터가 없는지 확인해보았다. 확인한 결과, 곡은 따로 발매되었지만 곡 제목과 가수 이름이 같은 데이터가 몇 개 존재했다(리패키지 앨범 등으로 같은 가수의 같은 곡이 여러 번 발매된 경우). 데이터를 수집할 때 key값으로 “곡 제목\_가수 이름”을 활용했기 때문에 별개의 곡에 대한 데이터가 하나의 데이터에 혼재되어버렸다. 이런 경우 정보가 혼합되어 학습에 활용할 수 없기 때문에 학습데이터에서 제외했다.  이상 데이터를 제거한 후, 분석에 활용할 feature extraction을 수행했다. 분석에 사용하기로 결정한 데이터의 feature은 시간별 평균 변동성이었다. 우선적으로 생각했던 feature은 시간별 평균 스트리밍 횟수였지만, 이를 사용하게 되면 곡마다 사용자수가 천차만별이기 때문에 스케일도 안 맞고 스트리밍 횟수가 많은 곡들이 변화폭도 클 것이기 때문에 이러한 부분을 고려하여 normalize된 변동성을 척도로 사용했다. 변동성은 “(현재 시각 평균 스트리밍 횟수 - 이전 시각 평균 스트리밍 횟수)/이전 시각 평균 스트리밍 횟수” 로 계산했다. 시간별 평균 변동성을 계산 하기 위해서는 실시간 스트리밍 횟수에 대한 정보가 필요했는데, 크롤링한 데이터에는 누적값에 대한 정보가 시간 순으로 들어있었다. 그래서 실시간 정보를 얻기 위해 인접한 데이터 간의 시간 차이를 계산하여(데이터에 날짜&시간 정보가 있었음) 시간 차이가 1시간이면 뒤 데이터의 누적 스트리밍 횟수에서 앞 데이터의 누적 스트리밍 횟수를 빼서 실시간 스트리밍 횟수를 계산했다. 시간별 스트리밍 횟수의 평균을 계산하기 위해 시간으로 groupby를 한 후 평균을 내주었고, 특정 시간에 대한 데이터가 없거나 모든 시간(0시부터 23시)에 대한 데이터 개수가 10개를 넘지 않으면 평균값의 안정성이 떨어진다고 판단하여 그 곡에 대한 데이터를 학습 데이터에서 제외하였다. 시간별 평균 스트리밍 횟수를 계산했으니 “(현재 시각 평균 스트리밍 횟수 - 이전 시각 평균 스트리밍 횟수)/이전 시각 평균 스트리밍 횟수”로 시간별 평균 변동성을 계산하였다. 곡 별로 이렇게 정제한 후 대략 750개 정도의 데이터를 얻을 수 있었고, 각 데이터는 ‘0시~1시’, ‘1시~2시’, … , ‘22시~24시’에 해당하는 변동성 정보를 갖고 있었기 때문에 23개의 열이 존재한다.  이 데이터는 feature만 있고 label이 없기 때문에 분류하기 위해 비지도 학습 방법인 K-Means Clustering과 Agglomerative Clustering을 활용했다. 두 가지 방법 모두 초모수인 K(분류할 그룹 개수)를 선택해줘야 하는데, 이를 위해 Silhoutte Plot과 Elbow Method를 활용하여 K=2로 선택했다. 애초의 목적이었던 ‘사재기 음원’, ‘팬덤형 음원’, ‘대중형 음원’을 분류하기 위해서는 K=3으로 선택해야 했지만, PCA를 활용해 데이터를 3차원으로 축소해서 시각화 해본 결과 K=2가 더 적합하다는 결론을 내렸다(이에 대한 설명은 뒤에서 더 자세히 다룰 예정). K=2로 설정한 채로 K-Means Clustering과 Agglomerative Clustering을 수행했고, 분류된 그룹 별로 데이터를 몇 개 샘플링하여 그래프로 그려본 결과 두 가지 방법 모두 ‘변동성이 상대적으로 큰 그룹’ vs. ‘변동성이 상대적으로 적은 그룹’으로 어느 정도 잘 분류한 것을 확인했다. |

1. **수행 결과 분석**

|  |
| --- |
| (주제 선정 당시 수행 목적대로 달성되었는지 기술)  주제 선정 당시 목적은 ‘사재기 음원’, ‘팬덤형 음원’, ‘대중형 음원’ 총 3개의 그룹으로 데이터를 분류하는 것이었다. 하지만 분석을 진행하면서 3개의 그룹으로 분류하는 것보다 2개의 그룹으로 분류하는 것이 더 적합하다는 것을 알게 되었다. 분석을 진행한 결과, 데이터가 ‘변동성이 상대적으로 큰 그룹’ vs. ‘변동성이 상대적으로 적은 그룹’으로 분류되었는데, ‘변동성이 상대적으로 적은 그룹’에 속한 곡과 가수들을 보면 다수가 아이돌(ex. NCT 127)이거나 트로트 가수(ex. 김호중)인 것을 확인할 수 있다. 이들은 대형 팬덤을 소유한 가수들로서, 팬들이 항시 스트리밍을 돌리고 있기 때문에 변동성이 상대적으로 적다고 추론해 볼 수 있다. 반면, ‘변동이 상대적으로 큰 그룹’에 속한 아이돌 가수(ex. 뉴진스)들도 있었는데, 이들은 큰 팬덤을 갖고 있으면서 동시에 대중성도 있어서 일반 대중들도 많이 듣기 때문에 앞서 언급한 가수들에 비해 변동성이 높다고 추론해볼 수 있다. |

1. **결론 및 느낀점**

|  |
| --- |
| (분석 결과 및 시사점 포함)  K-Means Clustering에서는 전체 데이터의 대략 35%가 ‘변동성이 상대적으로 적은 그룹(팬덤형 음원)’으로 분류되었고, Agglomerative Clustering에서는 전체 데이터의 대략 49%가 ‘변동성이 상대적으로 적은 그룹(팬덤형 음원)’으로 분류되었다. 분석 방법에 따라 어느 정도의 차이가 있기는 하지만, 현재 음원 시장에서 팬덤형 음원의 비율이 과도하게 높아 보인다. 음원을 듣지 않을 때도 본인이 응원하는 가수의 음원 성적을 위해 여러 기기에서 스트리밍을 켜두는 것(흔히 “스트리밍 공세”라고 부른다)은 음원 시장을 왜곡하고 다른 가수들의 기회를 빼앗을 수도 있다. 현재 이를 최대한 제지하기 위해 다수의 국내 음원 사이트에서 차트 집계 방식을 바꾸는 등 다양한 조치를 취하고 있지만 추가적인 조치가 필요해 보인다.  또한, 주제 선정 당시 분석 목적 중 하나였던 ‘사재기 음원 분류’를 달성하지 못한 것에 대해서도 이유를 고민해보았다. 처음 사재기 논란이 수면 위로 올라온 것은 2018년이었고, 논란이 극대화 되었던 것은 2019년이었다. 이 당시에는 사재기 의심 음원들이 새벽 시간에 갑자기 스트리밍 횟수가 급증하는 등 (대부분의 사람들이 취침을 취하고 있는 새벽시간은 보통 스트리밍 횟수가 떨어지는 것이 일반적이다) 상식적으로 이해하기 어려운 추이를 보였었다. 하지만 사재기 논란이 도마에 위에 오르고 사람들의 의심이 심해지자, 사재기 의심을 받는 음원들도 이 정도로 비정상적인 추이를 더 이상 보이지 않고 다른 음원들과 비슷한 추이를 보이게 되었다. 때문에, ‘사재기 음원 분류’를 하기 위해서는 2018~2019년도의 음원 데이터를 사용하는 것이 가장 이상적이라고 생각한다. 하지만 크롤링한 사이트에 이 기간의 데이터가 존재하지 않아 2020년 1월부터의 데이터를 사용했기 때문에 ‘사재기 음원 분류’를 하는데 어려움이 있었다고 추측해본다. |

1. **참고자료(APA 양식에 따른 작성)**

참고자료는 수업자료 외 없음

1. **팀원명 및 역할**

|  |  |  |
| --- | --- | --- |
| **학번** | **이름** | **역할** |
| 201715032 | 이하진 | 이하진 |
|  |  |  |
|  |  |  |
|  |  |  |