R 기반의 빅데이터 기술을 활용한 뉴스기사와 음원차트의 상관관계 분석†

하정철, 강동훈, 박재모, 길준민††
대구가톨릭대학교 IT공학부
e-mail: wjdcjf0219@gmail.com, kang11dh@naver.com, gt0849@naver.com, jmgil@cu.ac.kr

Correlation Analysis between News Articles and Music Charts using Big Data Technologies based on R

Jung-chul Ha, Dong-hoon Kang, Jae-mo Park, Joon-Min Gil
School of Information Technology Eng., Catholic University of Daegu.

요 약

빅데이터의 일종인 뉴스기사 중에 아이돌 그룹관련 뉴스기사는 아이돌 그룹의 대중적 인기에 힘입어 전체 연예계 기사 중에 점점 큰 비중을 차지하고 있다. 아이돌 그룹의 소속사는 여러 홍보 방법 중 뉴스기사의 노출을 통해 비교적 저렴한 비용으로 홍보하여 음원차트 순위 향상을 위해 노력하고 있다. 본 논문에서는 뉴스기사와 음원차트 간의 상관관계를 분석하여 뉴스기사의 노출이 효율적 홍보 수단인지를 알아보기 위해 먼저 감성분석을 통해 긍정기사와 부정기사가 음원차트 순위에 미치는 영향을 분석하고, 뉴스기사의 수가 많을수록 음원차트 순위가 상승하는지에 대해 알아보고자 한다. 이를 위해 본 논문에서는 R 언어를 이용하여 데이터 수집을 위한 웹 크롤러 설계, 회귀분석을 이용한 감성사전 구축 및 감성분석, 마지막으로 피어스만 상관계수를 이용한 상관관계 분석을 수행한다.

1. 서론

데이터의 양이 현재 테라바이트 급에서 향후에는 페타바이트 급, 엑사바이트 급도 넘어설 것으로 예상되는 가운데, 기존의 일반 기술로 수집, 저장, 관리, 분석하기 어려운 규모의 데이터를 빅데이터로 정의하고 있다[1]. 빅데이터 중 하나인 뉴스기사는 현실 세계에 일어나는 각종 현상에 대한 설명으로 미래의 정치, 경제, 사회, 기업 등과 관련하여 앞으로 어떤 변화가 발생할지에 대한 정보들을 포함하고 있다[2].

다양한 분야의 뉴스기사를 분석하고 의미를 추출하여 이를 활용하려는 많은 시도들이 활발하게 이루어지고 있다. 여러 분야의 뉴스기사 중에 아이돌 그룹의 뉴스기사는 아이돌 그룹의 대중적 인기에 힘입어 전체 연예계 뉴스기사 중에 점점 큰 비중을 차지하고 있다. 소속사에서는 뉴스기사를 통해 아이돌 그룹의 홍보, 특히 음원차트 순위 향상을 위해 이러한 홍보에 크게 의존하고 있다. 하지만, 아이돌 그룹과 관련되어 음원홍보, 스캔들, SNS 활동, 순위기록 등 다양한 유형의 뉴스들이 실시간으로 양산되고 있음에도 불구하고, 그 내용이 긍정적인지 부정적인지 명확히 파악하기가 쉽지 않다. 또한 뉴스가 다소 중립적인 뉘앙스로 음원시장의 긍정/부정 양쪽 의견을 모두 제시하는 경우가 많기 때문에 그 의미를 정확히 파악하는 것이 간단하지 않으며, 뉴스기사를 분석하는 사람의 주관에 따라 달라질 수 있다.

최근 아이돌 그룹의 대란 속에 소속사들이 과도한 비용을 지불하며 홍보에 매진하고 있다. 하지만 뉴스기사와 음원차트에 어떤 연관성을 갖고 있으며, 연관성이 존재한다면 그 연관성에 대한 분석은 아직 이루어지 않은 채 막연한 홍보에 치중하고 있다. 이에 본 논문에서는 최근 IT분야의 핵심기술인 빅데이터 기술을 이용하여 뉴스기사 수와 음원차트의 상관관계를 분석함으로써 그 연관성을 유도하고 이를 홍보비용 절감에 활용할 수 있음을 제시하고자 한다.

이를 위해 본 논문에서는 R 기반의 빅데이터 분석 기술 중 오피니언 마이닝을 이용하여 비정형 데이터인 뉴스기사에 대한 감성분석 및 뉴스 기사 내용에 대한 감성점수를 측정하여 부정기사가 음원차트에 끼치는 부정적인 영향력을 알아보고, 전체 기사 수와 음원차트 간 상관관계를 분석한다. 이를 위한 타겟으로 음원차트 1~100위권에 머무는 아이돌 그룹을 선정하여 분석한 자료를 바탕으로 현재 음원시장 상황과 홍보비용 측정에 관한 척도, 그리고 여러 소속사들의 홍보 트렌드를 제공하여 각 소속사들의 홍보비용 절감에 관한 대안책을 제시한다.

† 이 논문은 2014년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. NRF-2014R1A1A2055463).
†† 교신저자

본 논문의 구성은 다음과 같다. 2장에서는 하나의 가설과 연구 방법론 및 절차에 대해 설명하고, 3장에서는 데이터 수집을 위한 웹 크롤러, 4장에서는 감성사전 구축 및 감성분석, 5장에서는 기사 수와 음원차트 순위 간의 상관관계 분석, 마지막 6장에서 본 논문의 결론을 맺는다.

2. 연구 방법론

2.1 연구 가설 및 절차

연구에 앞서 다음과 같은 가설을 세워보았다.

[가설] 기사 내용이 부정일 경우 음원차트에 부정적 영향을 끼칠 것이다.

위 가설을 증명하기 위해 본 논문에서는 회귀분석을 통한 감성사전 구축 및 감성분석, 이를 통해 뉴스기사 수와 음원차트 간에 상관관계를 통계량 분석인 피어슨 상관계수를 이용하여 분석한다. 이러한 분석을 위해 본 논문의 연구는 첫째, 웹 크롤러 설계를 통한 데이터 수집, 둘째, 수집된 데이터에 대한 감성사전 구축 및 감성분석, 마지막으로 전체 뉴스기사 수와 음원차트 순위 간의 상관관계 분석의 순서로 수행되었다. 다음은 뉴스기사 수와 음원 차트 간의 상관관계 분석을 위해 본 논문에서 기본적으로 사용하는 통계적 방법을 간략히 기술한다.

2.2 회귀분석

회귀분석(regression analysis)은 한 변수가 다른 변수에 대해 미치는 영향을 추정할 수 있는 통계 기법이다. 회귀분석 방법에는 라쏘, 능형, 엘라스틱넷 등이 있으며, 이 중 엘라스틱넷이 라쏘와 능형의 장점을 동시에 만족하는 방법이므로 본 논문에서는 엘라스틱넷을 사용한다.

엘라스틱넷은 식 (1)처럼 라쏘 회귀분석의 λ_1 제약조건과 능형 회귀분석의 λ_2 제약조건을 결합한 방법이다. 식 (2)에서 $\alpha=0$이면 능형 회귀분석을, $\alpha=1$이면 라쏘 회귀분석을 나타내며, 엘라스틱넷은 0에서 1 사이의 α 값에 따른 라쏘 회귀분석과 능형 회귀분석의 볼록 결합(convex combination)에 해당한다.

$$\sum_{i=1}^{n}(y_i-\beta_0-\sum_{j=1}^{P}\beta_i x_{i.j})^2+\lambda_1\sum_{j=1}^{P}|\beta_j|+\lambda_2\sum_{j=1}^{P}|\beta_j|^2 \quad (1)$$

$$(1-\alpha)\sum_{j=1}^{P}|\beta_j|+\alpha\sum_{j=1}^{P}|\beta_j|^2 \le t \quad (2)$$

따라서 엘라스틱넷은 라쏘 회귀분석의 영향력이 없는 변수의 회귀 계수를 0으로 만들어 차원을 축소해 변수 선택이 가능한 장점과 능형 회귀분석의 전체적인 회귀 계수의 크기를 축소함으로써 관련성이 높은 설명 변수가 있을 때 변수들을 그룹화하여 다중공선성의 문제를 해결한 장점 두 가지를 동시에 만족하는 방법이다[3].

본 논문에서는 엘라스틱넷을 통해 나온 감성단어들이 연예 기사의 긍정, 부정을 판독하는데 의미있는 역할을 한다고 가정하여, 웹상에서 일반적으로 긍정 또는 부정으로 알려진 단어들과 엘라스틱넷을 사용한 결과 중에서 의미가 있다고 생각되는 상위 10개의 단어를 혼합하여 사전을 구축하였다. 구축된 사전을 바탕으로 긍정 단어는 +1점, 부정 단어는 -1점으로 점수를 매겨 단어의 포함여부에 따른 감성분석을 실시하였다.

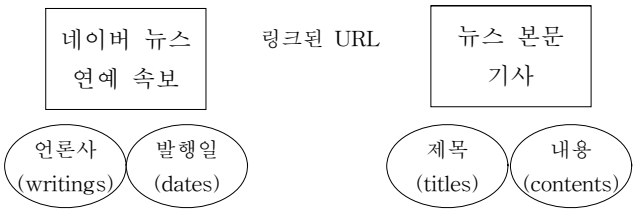
2.3 피어슨 상관계수

피어슨 상관계수(Pearson correlation coefficient)는 두 변수 간의 관련성을 얻기 위해 보편적으로 사용되는 방법이다. 즉, 두 변수 X와 Y가 함께 또는 따로 변하는 정도를 나타내는 개념으로서 X와 Y가 완전히 동일하면 +1, 전혀 다르면 0, 반대방향으로 완전히 동일하면 -1을 가진다. 일반적으로 다음과 같이 해석된다.

- r이 -1.0과 -0.7 사이이면, 강한 음적 선형관계,
- r이 -0.7과 -0.3 사이이면, 뚜렷한 음적 선형관계,
- r이 -0.3과 -0.1 사이이면, 약한 음적 선형관계,
- r이 -0.1과 +0.1 사이이면, 거의 무시될 수 있는 선형관계,
- r이 +0.1과 +0.3 사이이면, 약한 양적 선형관계,
- r이 +0.3과 +0.7 사이이면, 뚜렷한 양적 선형관계,
- r이 +0.7과 +1.0 사이이면, 강한 양적 선형관계

따라서 본 논문에서 사용하는 뉴스기사 수와 음원차트 순위의 경우 뉴스기사 수가 많을수록 그리고 음원차트 순위는 숫자가 낮을수록 높은 순위로 해석되기에 -1에 가까울수록 서로 간에 상관도가 높다고 볼 수 있다[4].

3. 데이터 수집을 위한 웹 크롤러



(그림 1) 웹 크롤러 구조

연예기사를 다루는 국내의 포털 사이트로는 네이버, 다음, 네이트 등 존재하며, 국내 사이트들의 특성상 비슷한 시간대에 비슷한 내용의 뉴스를 전달한다. 따라서 본 논문에서는 모든 포털 사이트가 아닌 국내 대표적 포털 사이트인 네이버 뉴스에서 한 페이지 당 20개의 기사를 제공하는 연예 속보 페이지(이하 뉴스 페이지)의 데이터를 수집하기 위한 웹 크롤러를 설계하였다. (그림 1)은 데이터 수집 과정에서 사용되는 웹 크롤러 구조를 보여주며, 본 논문의 웹 크롤러는 R에서 제공하는 Rvest 라이브러리[5]를 사용하였다. 이 웹 크롤러의 처리 과정은 첫번째 단계로 뉴스 페이지에서 언론사(writings), 발행일(dates), 그리고 사진에 링크된 URL을 데이터로 저장한다. 두번째 단계는 저장된 URL을 가지고 뉴스 본문 페이지로 넘어가 제목(titles), 내용(contents)을 수집한다. 날짜 별로 적게는 2천개, 많게는 8천개 이상의 뉴스를 내보내는데 본 논문의

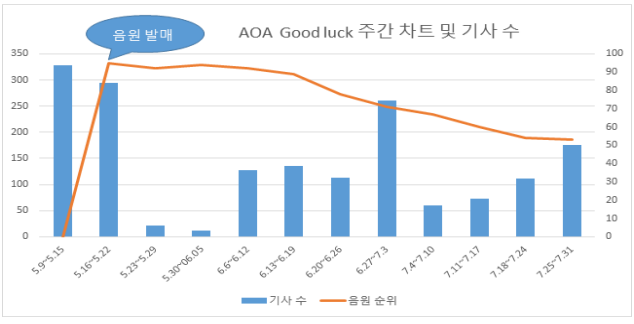
목표는 뉴스기사 수와 음원차트의 상관관계를 통한 트렌드 분석이기 때문에, 모든 뉴스기사를 수집하기 보다는 시간대가 고르게 분포된 각 날짜별 2천 건씩 일정량의 뉴스기사를 수집하도록 하였다. (그림 2)는 수집된 데이터의 일부를 보여주며, 이 그림과 같은 형태의 데이터를 월별 평균 6만 건씩 14년 07월부터 16년 07월까지 수집하여 약 150만 건에 해당하는 빅데이터 셋을 구축하였다.

titles	contents	writings	dates
'착하지 않은 여자들' 이두진 이하나에게 "엄마를 응서해 달…	사진 : KBS 2TV '착하지 않은 여자들'[헤럴드 리뷰스타=이…	헤럴드경제	2015-04-01 23:52
'크리임씬2' 하니 추리 계획터 '소녀 멘탈+메모 중독'	[TV리포트=윤미란 기자] '크리임씬2' 소녀 탐정, 하니의 추…	TV리포트	2015-04-01 23:51
'라디오스타' 이형도 "소속연예인 소유미, 꼭 뽑 것"	[TV리포트 = 이혜미 기자] 걔퍼이자 소속사대표인 이형도…	TV리포트	2015-04-01 23:51
'수요미식회' 김유석 "수타면으호 유명한 자장면집, 면이 달라"	[TV리포트=하수나 기자] 수타면으호 유명한 짜장면집은 어…	TV리포트	2015-04-01 23:51
'크리임씬2' 장진, 감독다운 예리한 관찰력+오지랖 웃음	[OSEN=김경주 기자] 장진 감독이 감독다운 예리한 관찰력…	OSEN	2015-04-01 23:51
안재욱 최현주 결혼, '포맨틱한 프로포즈' 세간의 부러움 사	안재욱 최현주 결혼 안재욱 최현주 결혼안재욱 최현주 결혼…	세계일보	2015-04-01 23:51
[포토] 정준호·이하정 부부, '다정하게 영화관 나들이~'	[변성현 기자] 배우 정준호·이하정 부부가 1일 오후 서울 영…	한국경제	2015-04-01 23:51
'수요미식회' 홍신애 "신성각 짜장면 먹고 울었다"	tvN'수요미식회' [텐아시아=조슬기 인턴기자] 홍신애가 수…	텐아시아	2015-04-01 23:50
김태우 김건과 계약 해지 "가족 아닌 차라리 나를 잘라해라"	[헤럴드경제]가수이자 소울상엔터테인먼트 대표인 김태우…	헤럴드경제	2015-04-01 23:50
김풍, ""무한도전·윤천소" 특집 작가한테 연락와 면접 봤었다"	라디오스타 (사진=방송캡처) '라디오스타' 김풍이 '무한도전…	한국경제	2015-04-01 23:50
'라디오스타' 이홍기, 한성호 대표 발언에 진짜 일할할 뻔 했다	라디오스타 이홍기 소속사 [텐아시아=오세림 인턴기자] '라…	텐아시아	2015-04-01 23:49
'투명인간' 정태호, 납규리 조련에 봉났다	[OSEN=이지영 기자] 납규리가 정태호를 조련(?)시키며 웃…	OSEN	2015-04-01 23:49
'엄마사람' 황혜영·이지현·현영의 리얼 육아…민낯 그대로 '…	'엄마사람' 황혜영·이지현·현영 /tvN 방송 캡처'엄마사람' 황…	한국경제	2015-04-01 23:49
[포토] 송재림, '귀여운 미소~'	[변성현 기자] 배우 송재림이 1일 오후 서울 영등포 타임스…	한국경제	2015-04-01 23:49

(그림 2) 수집된 데이터의 일부

4. 감성사전 구축 및 감성분석

앞서 기술한 [가설]을 검증하기 위해 수집된 뉴스기사 내용 중 긍정 또는 부정이라 생각되는 뉴스기사 각각 1천 건을 임의로 뽑은 후 이를 샘플 데이터로 사용하였다. 감성값(sentiment)은 내용을 읽어본 후 주관적인 판단 하에 긍정으로 간주된 문서는 1로, 부정으로 간주된 문서는 0으로 분류하였다. 기사 내용에 대한 형태소 분석의 경우 R 기반의 KoNLP 라이브러리[6]를 사용하여 명사, 동사, 형용사를 구분하였고 감성사전은 엘라스틱넷을 적용한 샘플 데이터의 회귀분석 결과와 일반 감성단어를 섞어서 구축하였다. 사전을 이용하여 전체적인 기사 데이터에 대한 감성분석을 실시한 결과 특정한 사건, 사고가 있지 않는 이상 대부분이 긍정기사라는 결과가 나왔다. 이에 따라 AOA의 리더 지민의 '긴또깡' 발언으로 인해 구설수에 오른 시기에 부정으로 분류된 기사 수와 음원 차트간의 상관관계를 분석하고자 '긴또깡' 사건이 발생한 5월 9일부터 신곡 Good Luck이 발매되고 2달 후인 7월 24일까지 약 16주간 데이터를 비교해 보았다.



(그림 3) 기사 수와 음원차트 그래프

(그림 3)에서 보는 것과 같이 파랑색 막대는 엘라스틱넷 감성사전을 통해 부정으로 판별된 기사 수를 주별로 보여주며, 주황색 실선은 음원차트 순위를 나타낸다. 이 그림에서 음원차트 순위의 경우 숫자가 낮을수록 높은 순위를 나타내기 때문에 원래 순위에서 -100을 더해준 후 절대값을 취하는 방식으로 음원차트 순위를 매겼다. (그림 3)에서 뉴스기사 수와 음원 차트 순위는 서로 간에 연관성이 없어 보이며, 이를 검증하기 위해서 R에서 제공하는 MASS 라이브러리[7]를 이용하여 피어스만 상관계수를 계산하였다.

```
> aoa1 = c(5,8,6,8,11,22,29,33,40,46,47)
> aoa2=c(295,22,12,128,136,113,260,60,72,112,176)
> aoa = data.frame(aoa1,aoa2)
> with(aoa, cor(x=aoa1,
+               y=aoa2,
+               use="complete.obs",
+               method=c("pearson")))
[1] 0.02848169
```

(그림 4) 피어스만 상관계수 결과

(그림 4)에서 볼 수 있듯이, 상관계수로 0.02848169라는 결과가 산출되었고 이는 두 요인이 거의 무시될 정도의 선형관계로 해석된다. 결과적으로 [가설]과는 달리 부정적 내용으로 판별된 기사라 할지라도 음원차트 순위에는 악영향을 끼치지 않는다. 따라서 이후의 상관관계 분석에서는 감성분석을 제외하고 전체 뉴스기사 수에 대한 음원차트 순위와의 상관관계 분석을 실시하도록 한다.

5. 기사 수와 음원차트 순위 간의 상관관계 분석

본 논문을 통해 비교·분석할 그룹은 아이돌 그룹으로서 아이돌 개인으로 활동하는 가수들을 제외한 그룹 위주로 기사와 음원차트 간의 상관관계 분석을 수행하고자 한다. 상관관계 분석에 앞서 전체 기사에서 아이돌 그룹관련 기사의 비율과 특정기간 동안 음원차트에서 1위부터 10위까지 해당하는 아이돌 그룹의 기사가 차지하는 비율을 살펴본다. <표 1>과 <표 2>는 각각 2015년 10월~2016년 3월과 2015년 4월~2015년 9월 기간의 음원차트와 기사순위를 보여준다. 이 표로부터 전체 기사 1,504,410건 중 아이돌 그룹관련 기사는 456,687건으로 약 30%에 해당하며, 이 기간 동안 음원차트에서 1위부터 10위까지 해당하는 그룹의 기사가 약 56%를 차지하고 있음을 알 수 있다. 또한, 연예계 3대 소속사 SM, YG, JYP에 속하는 가수들이 10위권 내 음원차트에서 항상 일정부분을 차지하고 있다는 것을 볼 수 있다. 하지만 음원차트 순위와는 다르게 뉴스기사 수는 SM 소속가수들이 많다는 것을 볼 수 있다. 아울러, 음원차트에서는 지속적으로 순위권에 있지만 뉴스기사 수는 정반대인 빅뱅(YG)과 엑소(SM)의 경우도 볼 수 있다.

<표 1>과 <표 2>에 나타난 전체 기사 수와 음원차트 순위를 살펴볼 때, 기사 수가 많다고 해서 음원차트 순위에 대한 영향력이 높다고 판단하기는 쉽지 않다. 그래서 이를 좀 더 분석해 보기 위해 <표 1>과 <표 2>의 조사기간 동안에 활동한 아이돌 그룹의 관련 기사와 음원차트 순위간의 상관관계 분석을 실시하였다. 음원이 출시된 후

첫 차트 진입 시기부터 3달(12주) 동안 주간 음원차트 순위와 기사 수를 비교·분석하여 상관계수를 산출하였다.

<표 3>은 음원차트 순위와 기사 수에 대한 상관관계를 보여주며, 이 그림의 결과는 강하거나 뚜렷한 상관관계를 의미하는 -1.0~-0.3 범위가 전체 중 57.1%를 차지하고 있음을 보여준다. 57.1%라는 수치로부터 기사 수와 음원차트 순위 사이에 연관성이 있다고 할 수 있다. 다시 말해, 아이돌 그룹의 기사의 수가 (긍정 기사나 부정 기사에 상관없이) 많으면 많을수록 음원차트 순위의 상승에 어느 정도의 영향을 끼칠 수 있지만 높은 수준으로 영향을 미친다고는 볼 수 없다. 이는 기사 수 이외에 다른 요인에 의해서도 음원차트 순위에 충분히 영향을 끼칠 수 있기 때문이다.

<표 1> 음원차트·기사 순위(2015년 10월~2016년 03월)

	음원차트 순위별 그룹	기사 순위별 그룹	기사 수	그룹별 비율(%)	전체기사 비율(%)
1	빅뱅	소녀시대(SM)	9761	14.5	8.2
2	엑소	EXID(바나나컬처)	8635	12.8	7.3
3	EXID	엑소(SM)	7769	11.5	6.6
4	레드벨벳	AOA(FNC)	7445	11.2	6.3
5	에이핑크	씨스타(스타쉽)	7101	10.5	6.0
6	AOA	레드벨벳(SM)	6918	10.3	5.8
7	걸스데이	빅뱅(YG)	6125	9.1	5.2
8	씨스타	에이핑크(플랜에이)	5411	8.0	4.6
9	소녀시대	마마무(RBW)	5006	7.4	4.2
10	마마무	걸스데이(드림티)	3179	4.7	2.7
합계					56.9

<표 2> 음원차트·기사 순위(2015년 04월~2015년 9월)

	음원차트 순위별 그룹	기사 순위별 그룹	기사 수	그룹별 비율(%)	전체기사 비율(%)
1	아이콘	소녀시대(SM)	8851	13.6	7.5
2	엑소	엑소(SM)	8836	13.5	7.5
3	빅뱅	AOA(FNC)	7852	12	6.7
4	마마무	EXID(바나나컬처)	7595	11.7	6.5
5	소녀시대	트와이스(JYP)	7228	10.1	6.2
6	AOA	레드벨벳(SM)	6350	9.7	5.4
7	EXID	아이콘(YG)	5934	9.0	5.1
8	러블리즈	러블리즈(울림)	4460	7.0	3.8
9	레드벨벳	빅뱅(YG)	4445	6.9	3.8
10	트와이스	마마무(RBW)	4177	6.5	3.6
합계					56.1

6. 결론

본 논문은 '홍보기사의 수가 많을수록 음원차트 순위가 높게 유지되는가?'에 가설에 대해서 단순히 기사 수와 음원차트 순위만 두고 보았을 때 '두 요인 사이에 연관성은 존재하지만 높은 수준은 아니다'라는 결론이 도출되었다. 이에 대한 원인은 음원차트 순위를 매기는데 있어서 기사 수 이외에 다른 요인들도 영향을 끼치기 때문이라 할 수 있다.

이러한 점으로 유추해 볼 때, 아이돌 그룹은 다양한 활동을 대중들에게 알리는 홍보가 중요하며, 이를 통해 대중성을 확보할 수 있다. YG가 상대적으로 적은 기사 수로

<표 3> 음원 순위·기사 수 상관계수와 누적비율

타이틀곡(가수)	상관계수	누적비율(%)
우리사랑하지말아요(빅뱅)	-1.0	4.8
아예(EXID)	-0.7	9.5
Remember(에이핑크)	-0.7	14.2
dumb dumb(레드벨벳)	-0.7	19.0
Sing for you(엑소)	-0.6	23.8
심쿵해(AOA)	-0.6	28.6
링마벨(걸스데이)	-0.6	33.3
shake it(씨스타)	-0.5	38.1
뱅뱅뱅(빅뱅)	-0.4	42.9
LOSER(빅뱅)	-0.4	47.6
음오아예(마마무)	-0.3	52.4
취향저격(아이콘)	-0.3	57.1
party(소녀시대)	-0.2	61.9
넌is뭔들(마마무)	-0.2	66.7
I Miss You(마마무)	-0.1	71.4
HOT PINK(EXID)	-0.1	76.2
lion heart(소녀시대)	0	81.0
Call me baby(엑소)	0.1	85.7
Love me right(엑소)	0.2	90.5
OOH-AHH하게(트와이스)	0.3	95.2
Ah-Choo(러블리즈)	0.5	100

높은 음원차트 순위를 유지하는 이유는 이러한 대중성을 잘 이끌어 내기 때문이라고 볼 수 있다. 신인 그룹이라면 음원 출시에 앞서 그룹명에 대한 홍보가 적절히 이루어져야 하며, 아무래도 생소한 그룹들이 음원을 출시한다면 음원이 좋다하더라도 곧 잘 잊혀 지기 때문이다.

향후 연구로는 뉴스기사 수 이외에 음원차트에 영향을 미치는 다른 요인을 추출하고 분석하는 것이다. 음원차트에 영향을 미치는 다른 요인들에 대해서 음원차트와의 상관관계를 분석한다면, 음원차트 순위에 가장 많은 영향을 끼치는 요인을 도출할 수 있을 것이며 이를 통해 좀 더 뚜렷한 대안을 제시할 수 있을 것이다.

참고문헌

[1] 김성재, 빅데이터의 충격, 한빛미디어, p. 26, 2013.
[2] 송치영, "뉴스가 금융시장에 미치는 영향에 관한 연구", 국제경제연구, 제8권, 제3호, 2005.
[3] Hui Zou and Trevor Hastie. "Regularization and variable selection via the elastic net", J. R. Statist. Soc. B., Vol. 67, No. 2, pp. 301-320, 2005.
[4] 김석우, 기초통계학, 학지사, pp. 96-97, 2007.
[5] https://cran.r-project.org/web/packages/rvest/, 2016.
[6] Heewon Jeon, KoNLP: Korean NLP Package, https://cran.r-project.org/web/packages/KoNLP/, 2016.
[7] https://cran.r-project.org/web/packages/MASS/, 2016.