닐로 날로잡기 프로젝트 동향분석서

작성일	‘18. 11. 3
작성자	남궁찬(품질 관리자)
참조논문	하정철, 강동훈, 박재모, 장으뜸, 이은영, 임성현, 길준민. (2016). 뉴스 기사와 음원차트 간의 상관관계 추출을 위한 R언어 기반 분석기의 설계 및 구현. 한국정보과학회 학술발표논문집, , 122-124.
결과도출방법	크롤링을 통해 직접 데이터를 수집하고 R 기반의 빅데이터 분석 기술 중 하나인 오피니언 마이닝을 적용하여 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.
사용기술	설명
	적용가능여부
회귀분석	한 변수가 다른 변수에 대해 미치는 영향을 추정할 수 있는 통계 기법, 라쏘, 능형(릿지), 엘라스틱넷 등이 있다. 해당 논문에서는 엘라스틱넷 기술을 사용하여 감성분석을 위한 감성사전을 구축함.
	해당 프로젝트에서는 많은 변수를 사용하지 않으므로 릿지 회귀 방법을 적용하는 것이 적절해 보임. Python Scikit-Learn 라이브러리를 활용.
피어슨 상관계수	두 변수간의 관련성을 얻기 위해 보편적으로 사용되는 방법, 상관계수가 -1.0에 가까우면 강한 음적 선형관계, +1.0에 가까우면 강한 양적 선형관계를 가진다. 해당 논문에서는 뉴스기사 수와 음원차트 순위의 관계를 비교한다. 뉴스기사 수가 많고 음원순위의 숫자가 낮을수록 -1.0에 가까워지며 상관도가 높다고 볼 수 있다.
	변수간의 관련성을 얻기 위해 사용하기 적절해 보임. pandas 라이브러리에서 지원하고 있음
오피니언 마이닝 (감성분석)	텍스트에 나타난 사람들의 의견이나 성향 같은 주관적인 데이터를 분석하는 자연어 처리 기술. 해당 논문에서는 뉴스 기사의 텍스트를 분석하여 긍정적인 뉴스인지 부적적인 뉴스인지 여부를 판별하기 위해 사용함.
	화제성 변수의 긍정적, 부정적 여부는 음원순위에 영향을 끼치지 않을 것으로 보임. (ex. 노이즈 마케팅) 따라서 적용하기에 부적절.

작성일	‘18. 11. 14	작성자	남궁찬(품질 관리자)
참조논문	Eva Zangerle, Martin Pichl, Benedikt Hupfauf, G¨unther Specht.(2016). Can microblogs predict music charts? An analysis of the relationship between #Nowplaying Tweets and Music Charts		
연구 주제	#Nowplaying 트윗이 빌보트 차트와 어느정도 유사한가		
	#Nowplaying 트윗이 빌보트 차트와 시간적으로 어떤 관련이 있는가		
	트위터 데이터가 뮤직 차트를 예측하는 데 사용될 수 있는가		
결과도출방법	관련 연구에서는 트위터에서의 음원 유명세, 아티스트 유명세, 음원이 빌보트 탑 100에 진입한 주 수 의 세 가지 속성을 사용했다. 특정 음원의 차트 랭킹과 세 속성의 피어슨 상관관계 분석했다. 음원 유명세가 가장 높은 상관관계를 보였다. 순위를 예측하기 위해 선형회귀, 이차선형회귀, SVR회귀 모델을 사용하였고, SVR회귀 모델이 가장 좋은 성능을 보였다. 히트 예측을 위해서는 데이터를 히트와 그렇지 않은 데이터로 나누고 랜덤 포레스트 분류를 사용하였다. 본 연구에서는 첫 번째로 랭킹의 상관관계를 알아보기 위해 트랙의 플레이 수 로그, 아티스트의 플레이 수 로그, 노래가 빌보드 탑100에 진입했던 주 수를 상관관계 분석했다. 두 번째로 트윗과 차트의 시간적 관계를 알아보기 위해 두 변수를 교차 상관분석했다. 세 번째로 빌보드 차트를 예측한다.		
사용기술	설명		
	적용가능여부		
Spearman 계수	상관 관계를 분석하고자 하는 변수의 분포가 심각하게 정규분포를 벗어난다거나 또는 두 변수가 **순위 척도** 자료일 때 사용하는 값,		
	해당 프로젝트에서는 네이버 트렌드와 음원차트 두 가지 변수의 상관관계를 분석한다. 두 변수는 순위 척도를 나타내므로 Pearson보다 Spearman 계수를 적용하는 것이 더 적절해 보인다.		
교차 상관분석	두 변수간의 시간차가 있는 경우 상관성이 낮게 나오는데, 이 시간차를 해결하기 위한 상관분석		
	네이버 트렌드와 음원차트 변수간의 시간차에 따른 관련성을 얻기 위해 활용할 수 있을 것으로 보임. numpy 라이브러리에서 지원함.		