

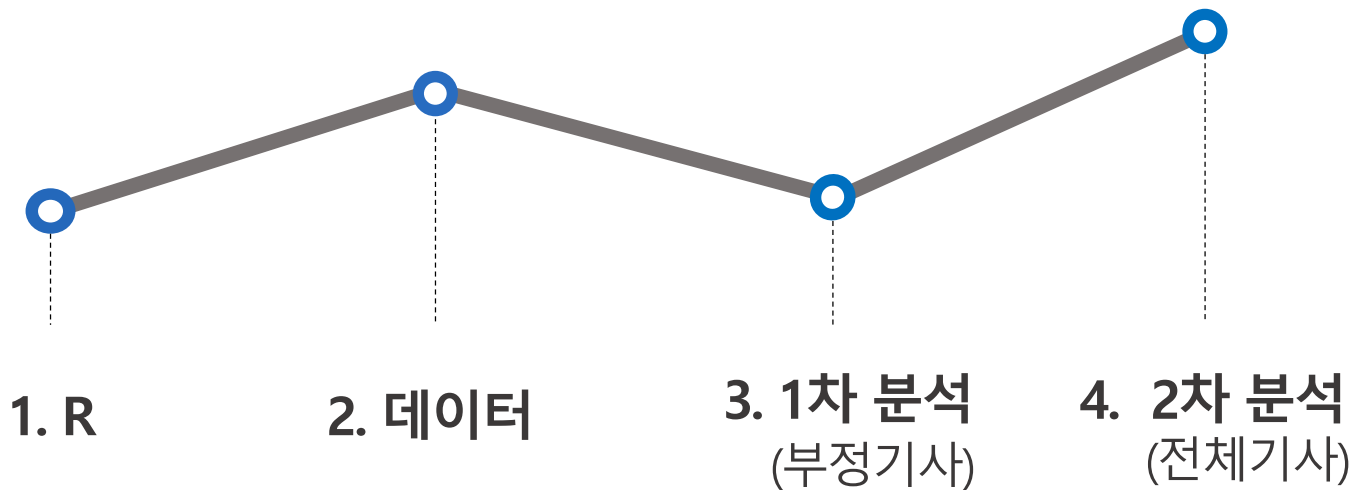


아이돌그룹 연예기사 개수와 음원차트 상관관계

하정철

2016.10.14
데이터야 놀자

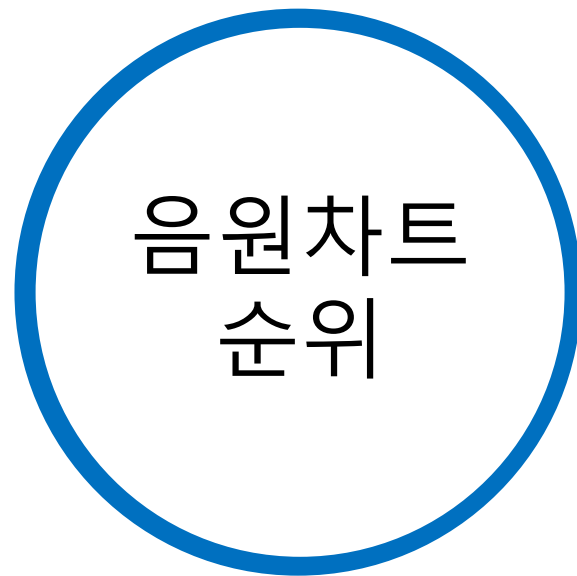
INDEX



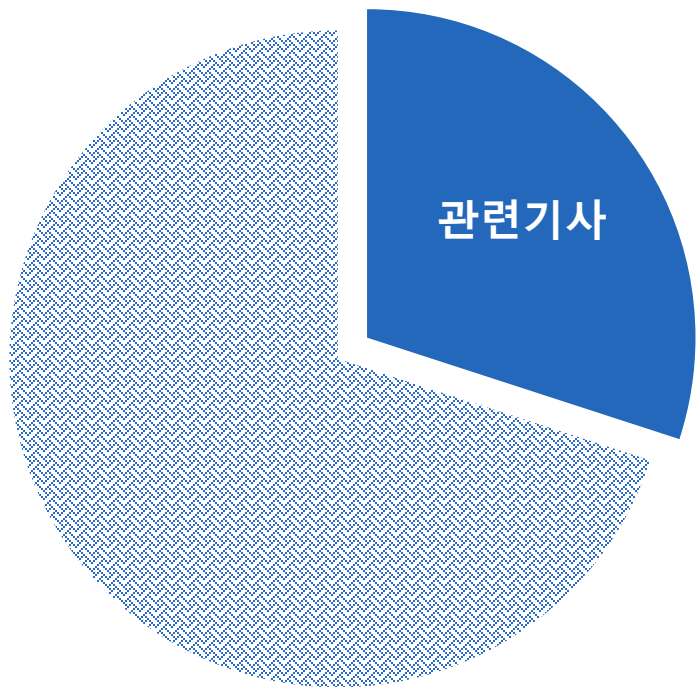




상관관계



아이돌 가수 관련 기사의 비율



전체 기사 : 1,504,410
관련 기사 : 456,687

약 30%

1. R



통계 (statistical computing)

데이터 마이닝 (data mining)

그래픽 (graphics)

2. 데이터 수집

웹 크롤러(web crawler)

웹 페이지를 방문해, 각종 정보를 자동적으로 수집하는 프로그램

Rvest

패키지

`read_html()`

`html_nodes()`

`html_attr()`

`html_text()`

Rvest를 이용해 **NAVER** 뉴스 속보 중에서

연예 부분을 수집해보자!

요약형

NAVER 뉴스
연예 속보

언론사
writings

발행일
dates

본문 URL

요약형

NAVER 뉴스
연예 속보

언론사
writings

발행일
dates

본문 URL

제목
titles

내용
contents

read_html(URL) = html 파싱

요약형

NAVER 뉴스
연예 속보

언론사
writings

발행일
dates

본문 기사

제목
titles

내용
contents

요약형

NAVER 뉴스
연예 속보

언론사
writings

발행일
dates

html_nodes()

html_attr()

메타정보

본문 기사

제목
titles

내용
contents

요약형

NAVER 뉴스 연예 속보

페이지당 20건
하루 2000건

다혜선 장근석과 아프로디테 ... TV리포트 | 2016-10-06 23:51

[포인트1분] '엄마가뒤통자' 최민수, 강주은 향해 "바가지다"

[헤럴드POP=이진숙 기자] 최민수가 가족을 위한 아침상을 준비했다.6일 TV조선을 통해 방송된 '엄
마가 뒤통자' ... 헤럴드POP | 2016-10-06 23:50

'해피투게더' 예원 "그 사건 후 일이 없어 맞혀지는 건가 싶어"

'해피투게더'의 예원이 이태임과의 '한말 논란' 이후 첫 지상파 출연을 한 소감을 밝혔다. 6일
오후 방송된 KBS2 '해피투게 ... MBN | 2016-10-06 23:50

'해투3' 홍진영 "가수 겸 배우가 최근 대서, 난 거리두고 있다"

[마이데일리 = 이승길 기자] 가수 홍진영이 최근 연예인에게 대서를 받은 사실을 고백했다.6일 방
송된 KBS 2TV ... 마이데일리 | 2016-10-06 23:50

1 2 3 4 5 6 7 8 9 10 | 다음 >

10월8일(토) · 10월7일(금) · 10월6일(목) · 10월5일(수) · 10월4일(화)

< 이전페이지로 돌아가기 | 맨위로

- 3 박지일 윤시운 NEW
- 4 설악산 첫 얼음 NEW
- 5 송해교 익플러 벌금형 선고
- 6 해경 고속단정
- 7 경찰서서 30대 분신 NEW
- 8 그것이 알고싶다 대구 희망원
- 9 내년부터 카드대금 연체 NEW
- 10 두산 장원준 NEW

2016.10.09 17:00 기준

본문 URL



본문 기사로 이동



백승주, 김제동 영창 발언 진상규명 요구...국감 증인 채택될까?

'김제동 영창' 발언에 백승주 진상규명 요구방송인 김제동(42)이 과거 방송 프로그램에 출연해 군 복무 시절을 회상한 발언이 뒤 ... 스포츠경향 | 2016-10-06 23:59

언론사

발행일

기사제목

스포츠경향 Sports KyungHwang

글꼴 - +

백승주, 김제동 영창 발언 진상규명 요구...국감 증인 채택될까?

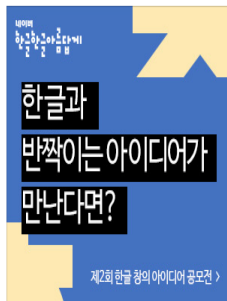
기사입력 2016.10.06 오후 11:59 | 기사원문 | 댓글 38

'김제동 영창' 발언에 백승주 진상규명 요구

방송인 김제동(42)이 과거 방송 프로그램에 출연해 군 복무 시절을 회상한 발언이 뒤늦게 논란이 되고 있다.



사진 김제동 페이스북



많이 본 뉴스 TV연예 종합 더보기

- 1 [BFF@곡성] '악마의 입담'...쿠니무라 준 밝힌 #나눔진..
- 2 '복면가왕' 지도 정제는 김국환...황진미리연 3R 진출
- 3 응답하라 신원호 PD "내년 자기작 준비 중"
- 4 [X연장] '시든 나온다'...김원석 감독이 밝힌 '미생' 그..
- 5 '색선' 주인공 '장동건보다 내가 더 잘생겼다'
- 6 '슈탈' 보검이보다 대박이 빠지면 줄구없는 마성배이비..
- 7 '마리탈' 측 '손연재 MLT-36 출연, 우주소년 성소와 리..
- 8 [BFF중합] '김민희는고양이' 사라진 아가씨 빈자리 채운..
- 9 '복면가왕' 팝콘소년 세 가왕 동극...'에라리디오는 정동..
- 10 POP이슈'악 좀 달라 달라'...소문난 전지 이수라의 말..

[V LIVE] V 생중계



< 10/12(수) 8:30PM NCT DREAM의 탐구생활 >














[1] "http://news.naver.com/main/list.nhn?sid1=106&mid=sec&mode=LSD&date=20160201%20page%3D%20&page=4"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=312&aid=0000170007"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=311&aid=0000572890"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=082&aid=0000572862"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=213&aid=0000841074"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=112&aid=0002773117"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=468&aid=0000109565"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=117&aid=0002720572"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=382&aid=0000446450"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=112&aid=0002773116"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=112&aid=0002773115"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=468&aid=0000109564"
[1] "http://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=106&oid=076&aid=0002884698"

data.frame으로 저장 한 후, CSV 파일로 관리

	titles	contents	writings	dates
1	'문제적남자' 지코 "박경과 인기 라이벌? 박경보단 내가 많았...	[엑스포츠뉴스=전아람 기자] 그룹 블랙비 지코가 박경과의 ...	엑스포츠뉴스	2016-05-01 23:57
2	꽃보다 예쁜 정다빈, '촬영 인증샷 공개'	[SBS funE 연예뉴스팀]꽃보다 예쁜 정다빈, '촬영 인증샷 공...	SBS funE	2016-05-01 23:55
3	'문제적남자' 지코, 학창시절 고백 "박경 내게 잘 보였어야 했...	[스타뉴스 김소희 인턴기자] /사진=tvN '문제적 남자' 방송...	스타뉴스	2016-05-01 23:54
4	황승언, '독면가왕' 출연소감 고백 "부들부들 떨었다"	사진:황승언 인스타그램[헤럴드POP=강수정 기자] 영화배우...	헤럴드POP	2016-05-01 23:53
5	방탄소년단, 최종 버닝맨의 주인공은?(종합)	[헤럴드POP=박세영 기자]방탄소년단 버닝맨의 주인공으로 ...	헤럴드POP	2016-05-01 23:53
6	'문제적남자' 지코 "박경은 전교, 난 학교 후문에서 놀았다"	[엑스포츠뉴스=전아람 기자] 그룹 블랙비 지코가 박경의 천...	엑스포츠뉴스	2016-05-01 23:53
7	"갈수록 훈훈하다" 류준열, 훗칠한 매력 보이며 '미소 만개'	[한국경제TV 트렌드연예팀 조은애 기자] 배우 류준열의 훈...	한국경제TV	2016-05-01 23:51
8	'런닝맨' 김지원, 오란씨걸 재연 "하늘에서 별을 따라~"	(서울=뉴스1 스타) 이진욱 기자 = '런닝맨' 김지원이 오란씨...	뉴스1	2016-05-01 23:50
9	정다빈, 옥중화 본방사수 목려..."우려했던 이목구비"	[SBS funE 연예뉴스팀]정다빈, 옥중화 본방사수 목려..."우...	SBS funE	2016-05-01 23:50
10	방탄소년단 렘몬스터 "청춘에 대해 고민...그 순간이 화양연...	[헤럴드POP=박세영 기자]방탄소년단 렘몬스터가 신곡을 준...	헤럴드POP	2016-05-01 23:49

14년 7월 ~ 16년 7월

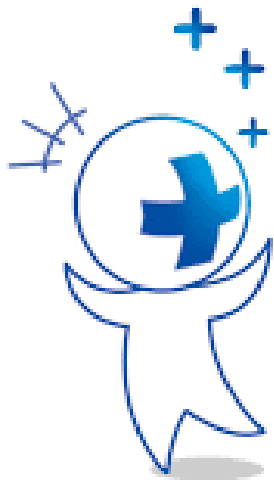
약 150만건의 기사 데이터

이름	수정한 날짜	유형	크기
 result(1407)	2016-08-23 오후 ...	Microsoft Excel 심표로 구분된 값 파일	67,932KB
 result(1408)	2016-08-23 오후 ...	Microsoft Excel 심표로 구분된 값 파일	64,329KB
 result(1409)	2016-08-23 오후 ...	Microsoft Excel 심표로 구분된 값 파일	67,765KB
 result(1410)	2016-08-23 오후 ...	Microsoft Excel 심표로 구분된 값 파일	66,745KB
 result(1411)	2016-08-23 오후 ...	Microsoft Excel 심표로 구분된 값 파일	66,286KB
 result(1412)	2016-08-23 오후 ...	Microsoft Excel 심표로 구분된 값 파일	67,147KB
		●	
		●	
		●	
 result(1601)	2016-08-24 오전 ...	Microsoft Excel 심표로 구분된 값 파일	69,048KB
 result(1602)	2016-08-28 오전 ...	Microsoft Excel 심표로 구분된 값 파일	60,145KB
 result(1603)	2016-08-26 오전 ...	Microsoft Excel 심표로 구분된 값 파일	66,931KB
 result(1604)	2016-08-26 오후 ...	Microsoft Excel 심표로 구분된 값 파일	64,237KB
 result(1605)	2016-08-26 오전 ...	Microsoft Excel 심표로 구분된 값 파일	65,469KB
 result(1606)	2016-09-14 오후 ...	Microsoft Excel 심표로 구분된 값 파일	65,425KB
 result(1607)	2016-09-11 오후 ...	Microsoft Excel 심표로 구분된 값 파일	67,428KB

3. 1차 분석(부정기사)

if

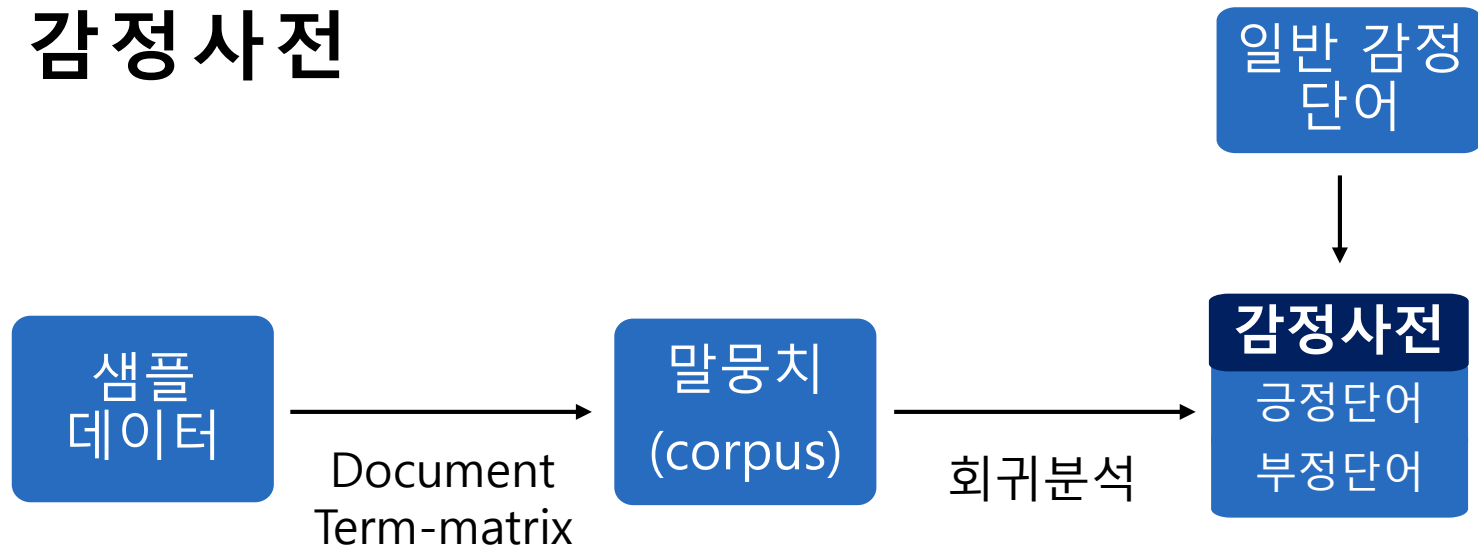
긍정 기사



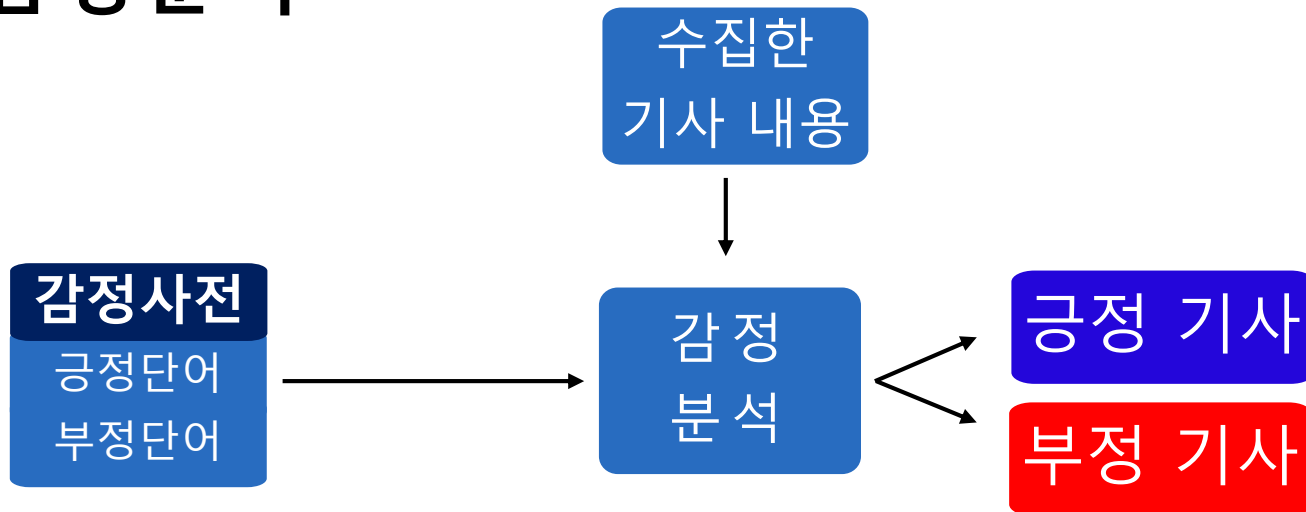
부정 기사



1. 감정사전



2. 감정 분석



부정 기사

상관관계 분석



샘플
데이터



말뭉치
(corpus)



감정사전



긍정/부정 각 1천건

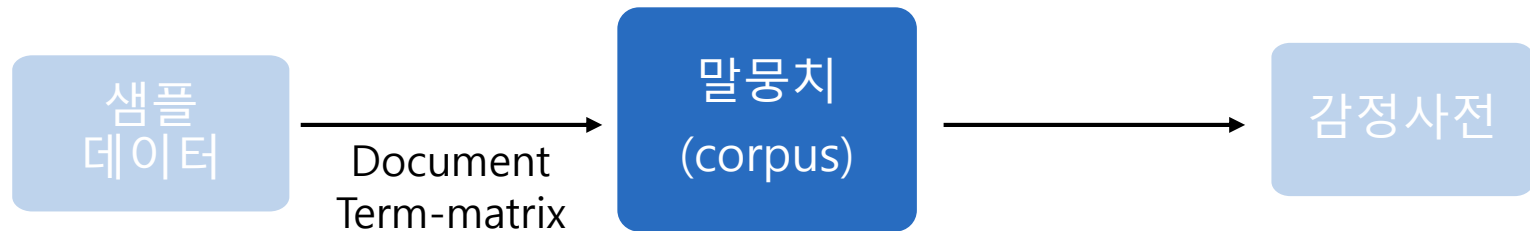
감정값(sentiment)

긍정 1

부정 0



회귀분석에
쓰여요!



특정 목적을 가지고 추출한 언어의 집합

감정에 영향이 없는 구두점, 숫자, 단어

Document-Term matrix (tm)

샘플
데이터

Document
Term-matrix

말뭉치
(corpus)

→

감정사전

한글 단어 분류 How?



KoNLP 패키지

명사, 형용사, 동사 구분

샘플
데이터

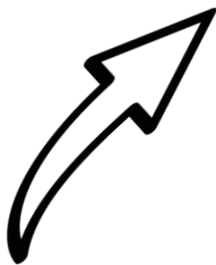
Document
Term-matrix

말뭉치
(corpus)

→

감정사전

```
cps <- Corpus(VectorSource(boa$contents)) #tm 자료구조 corpus로 만들어  
dtm <- DocumentTermMatrix(cps,  
  control=list(tokenize=ko.words, # 단어를 쪼개고  
              removePunctuation=T, # 구두점  
              removeNumbers=T, # 숫자  
              wordLengths=c(2, 5), # 단어 길이  
              weighting=weightT))
```



다양한 옵션



회귀분석 (**glmnet**)

단어가 사용 유무 정도로 회귀분석을 실시

라쏘(lasso) : 작은 회귀계수를 0으로 만듦

릿지(ridge) : 전반적으로 회귀계수를 줄여줌

엘라스틱넷(elastic net) : 라쏘 + 릿지


```

X <- as.matrix(dtm)
Y <- sample$sentiment

res.elastic <- cv.glmnet(X, Y, family = "binomial", alpha = .5,
                          nfolds = 4, type.measure="class")

coef.elastic <- coef(res.elastic, s="lambda.min")[,1]

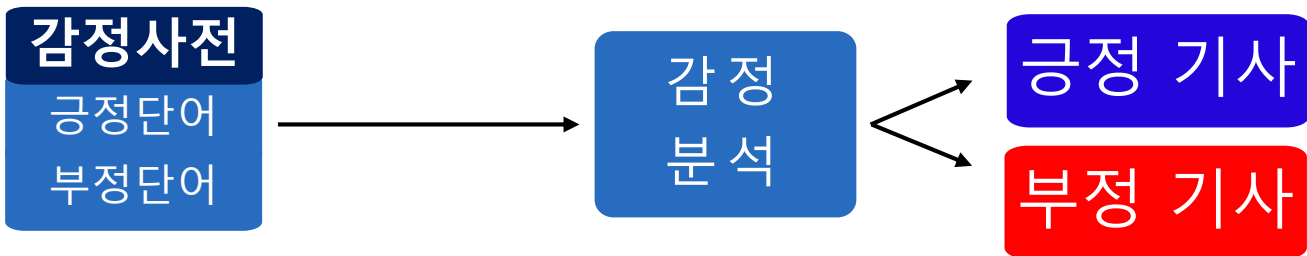
```

```
> pos.elastic[1:20]
```

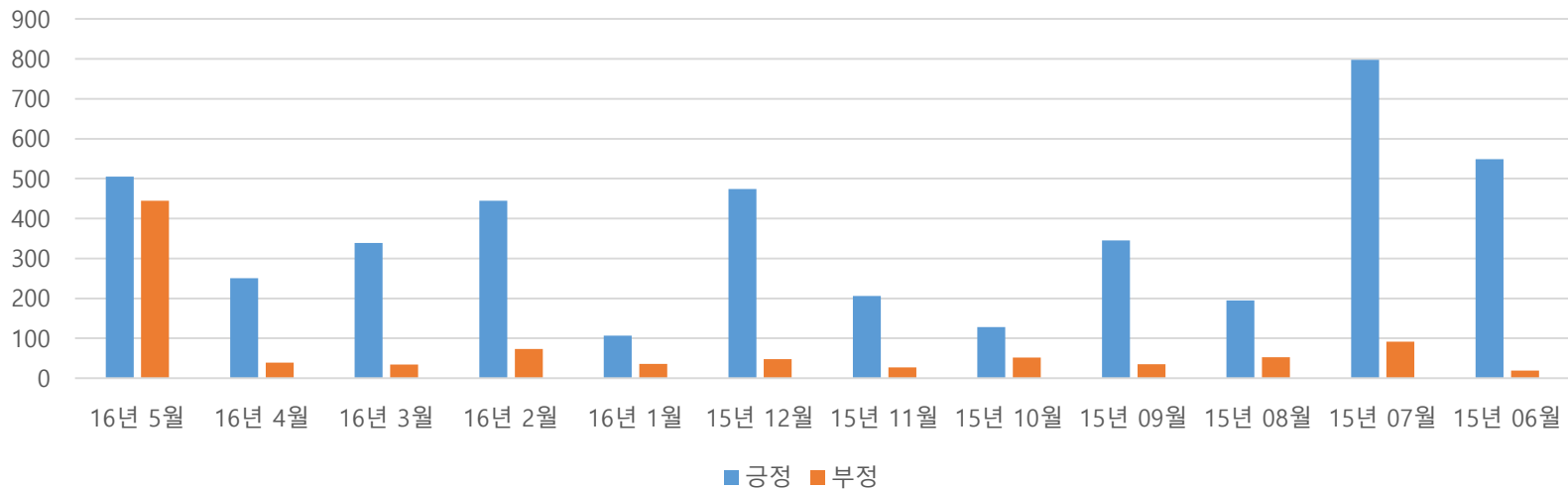
걸그룹	무대	모바일	웬디	공개	선보이	레드벨벳
44.520701	18.225396	11.727538	11.556724	10.919651	10.442568	10.311873
매력	그룹	데뷔	뉴스스탠드	마마무	눈길	최고
9.622285	8.853587	6.647353	6.440536	5.621460	5.360958	5.302264
멤버들	논란에도	상위권	준비	마마무가	상암동	
4.987219	4.611096	4.179403	3.870763	3.744790	3.694436	

```
> neg.elastic[1:20]
```

논란	성폭행	의사	떠올리	사건	밝히	사실
-18.611747	-13.618654	-11.243598	-11.230186	-9.960996	-9.174187	-8.815463
안중근	국내최대	혐의	대하	경찰	역사	서울경제섬
-8.638332	-8.116058	-7.329697	-7.135032	-6.617053	-6.468732	-6.236040
클릭클릭	당시	배우	자기	못하	연합뉴스	
-6.018028	-6.016456	-5.005866	-4.524220	-4.274470	-4.179396	



긍정 / 부정 기사 개수





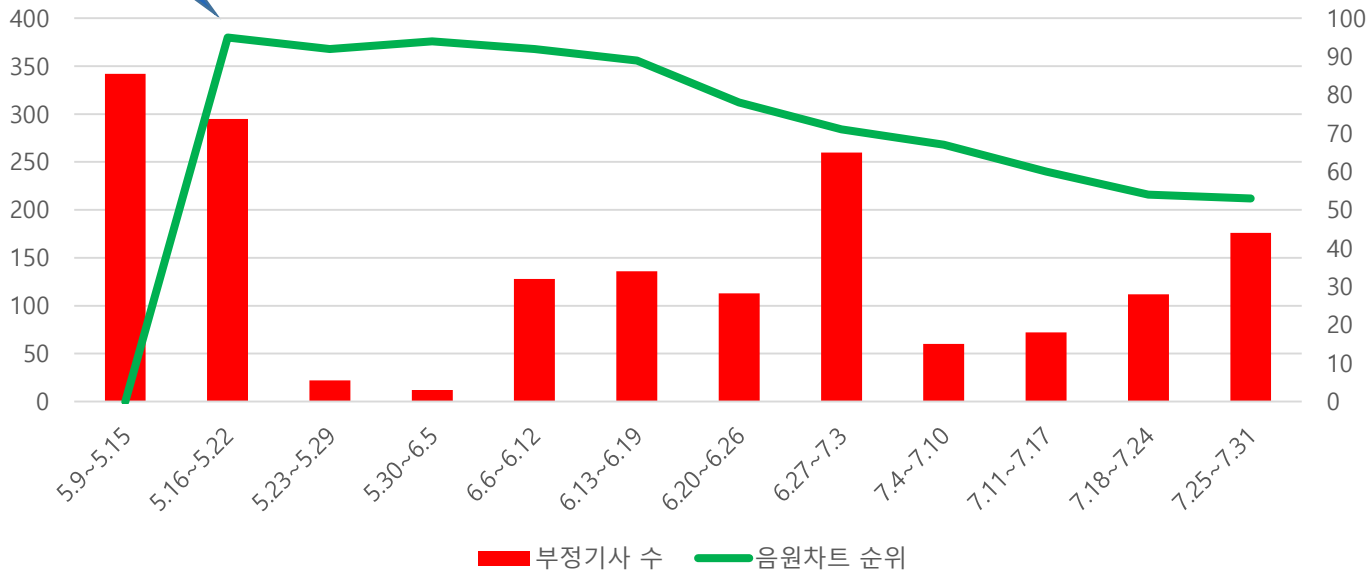
5월 9일, AOA



5월 9일, AOA '긴또깡' 사건

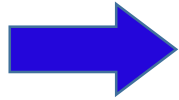
차트
진입

Good Luck



• 피어슨 상관계수

두 변수 간의 관련성을 얻기 위한 방법
즉, 두 변수 X와 Y가 함께 또는 따로 변하는 정도



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

MASS 패키지를 사용한 상관계수

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

1차 결론

상관계수 0.02848169



[가정]과는 부정기사와 음원차트 순위간의 상관도는 미비하다

4. 2차 분석(전체기사)

아이돌 전체 기사와 1~10위 그룹 관련 기사의 비율

순위					
	음원차트	기사	기사 수	그룹별 비율(%)	전체 기사 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
6	AOA	레드벨벳(SM)	6918	10.3	5.8
7	걸스데이	빅뱅(YG)	6125	9.1	5.2
8	씨스타	에이핑크 (플랜에이)	5411	8	4.6
9	소녀시대	마마무(RBW)	5006	7.4	4.2
10	마마무	걸스데이 (드림티)	3179	4.7	2.7
기타					56.9

15년 04월 ~ 15년 09월

순위					
	음원차트	기사	기사 수	그룹별 비율(%)	전체 기사 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	5.1
8	러블리즈	러블리즈 (울림)	4460	7	3.8
9	레드벨벳	빅뱅(YG)	4445	6.9	3.8
10	트와이스	마마무(RBW)	4177	6.5	3.6
기타					56.1










15년 10월 ~ 16년 03월

순위권 그룹 타이틀곡(21곡) 주간순위

ice cream cake	3.16	16	13	12	11	14	15	17	17	18	22
call me babay	3.23	26	6	6	9	9	11	13	14	19	21
LOSER	4.27	3	1	1	3	2	3	5	6	10	10
love me right	6.1	6	4	5	8	11	19	24	29	38	46
뱅뱅뱅	6.1	1	1	2	2	6	9	9	8	7	10
음오아예	6.15	38	5	7	11	14	13	12	14	16	26
IF you	6.29	2	2	8	10	13	15	17	29	32	34
우리 사랑하지 말아요	8.3	2	1	4	11	11	14	16	21	22	28
party	7.6	3	5	9	11	18	21	37	38	39	42
lion heart	8.17	23	17	9	5	7	12	14	19	23	30
dumb dumb	9.14	6	7	9	12	14	25	27	29	27	29
Ah-Choo	9.28	83	65	62	64	62	64	56	48	29	35
OOH-AHH하게	10.2	55	31	21	13	13	11	17	12	12	11
sing for you	12.1	9	5	6	9	12	12	16	24	33	34
I Miss You	1.25	19	10	11	11	11	10	15	16	24	34
넌 is 원들	2.22	8	1	1	1	1	2	4	4	5	7

이름

^

-  10대 그룹 음원순위.상관계수
-  aoa 주간순위
-  apink 주간순위
-  exid 주간순위
-  걸스데이 주간순위
-  씨스타 주간순위
-  아이콘 주간순위
-  원더걸스 주간순위
-  워너 주간순위

타이틀곡 첫 차트 진입 후 기사 개수

2015-03-09	ice cream cake		
2015-03-16	1616	call me baby	
2015-03-23	1613	2145	
2015-03-30	3607	1600	
2015-04-06	2406	931	
2015-04-13	1032	1040	
2015-04-20	1623	1149	LOSER
2015-04-27	1480	306	1255
2015-05-04	2602	434	2641
2015-05-11	311	2757	1874
2015-05-18	3807	792	2372
2015-05-25	366	752	802
2015-06-01	1455	448	3321
2015-06-08		1346	1627
2015-06-15			1802
2015-06-22			1273
2015-06-29			1908
2015-07-06			880
2015-07-13			646

2015-09-07		dumb dumb		
2015-09-14		4587		
2015-09-21		1813	ah-choo	
2015-09-28		3087	1108	
2015-10-05		4053	1078	
2015-10-12		1781	2195	ooh-ahh
2015-10-19		612	1308	3624
2015-10-26		1334	1700	2009
2015-11-02		2685	562	1433
2015-11-09		670	2613	2334
2015-11-16		633	201	1472
2015-11-23		799	159	1992
2015-11-30	sing for you	1545	138	1137
2015-12-07	590		979	539
2015-12-14	912		1764	530
2015-12-21	1940			4364
2015-12-28	1927			2435
2016-01-04	1030			1577

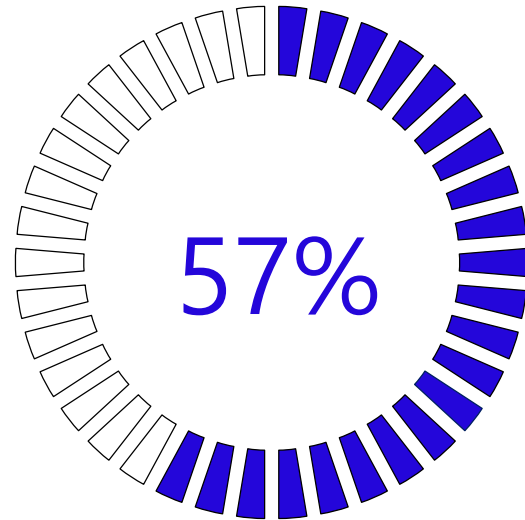
순위권 그룹 타이틀곡과 음원차트 주간순위 상관계수

	타이틀곡(가수)	상관계수
1	우리사랑하지 말아요(빅뱅)	-1.0
2	아예(EXID)	-0.7
3	Remember (에이핑크)	-0.7
4	dumb dumb (레드벨벳)	-0.7
5	Sing for you(엑소)	-0.6
6	심쿵해(AOA)	-0.6
7	링마벨(걸스데이)	-0.6
8	shake it(씨스타)	-0.5
9	뱅뱅뱅(빅뱅)	-0.4
10	LOSER(빅뱅)	-0.4
11	음오아예(마마무)	-0.3
12	취향저격(아이콘)	-0.3

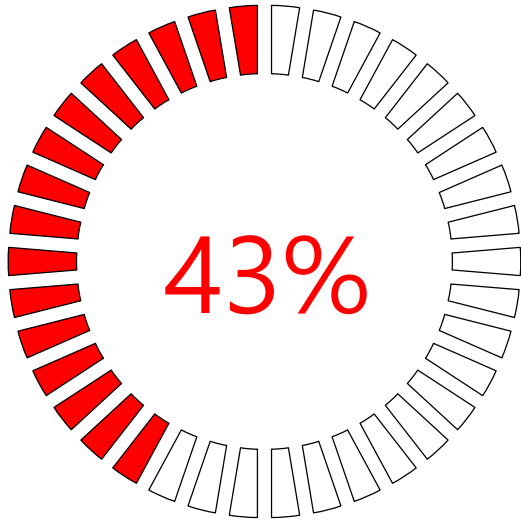
	타이틀곡(가수)	상관계수
1	party(소녀시대)	-0.2
2	넌is원들(마마무)	-0.2
3	I Miss You(마마무)	-0.1
4	HOT PINK(EXID)	-0.1
5	lion heart(소녀시대)	0
6	Call me baby(엑소)	0.1
7	Love me right(엑소)	0.2
8	OOH-AHH하게 (트와이스)	0.3
9	Ah-Choo(러블리즈)	0.5

순위권 그룹 타이틀곡과 음원차트 주간순위 상관계수

	타이틀곡(가수)	상관계수
1	우리사랑하지 말아요(빅뱅)	-1.0
2	아예(EXID)	-0.7
3	Remember (에이핑크)	-0.7
4	dumb dumb (레드벨벳)	-0.7
5	Sing for you(엑소)	-0.6
6	심쿵해(AOA)	-0.6
7	링마벨(걸스데이)	-0.6
8	shake it(씨스타)	-0.5
9	뱅뱅뱅(빅뱅)	-0.4
10	LOSER(빅뱅)	-0.4
11	음오아예(마마무)	-0.3
12	취향저격(아이콘)	-0.3

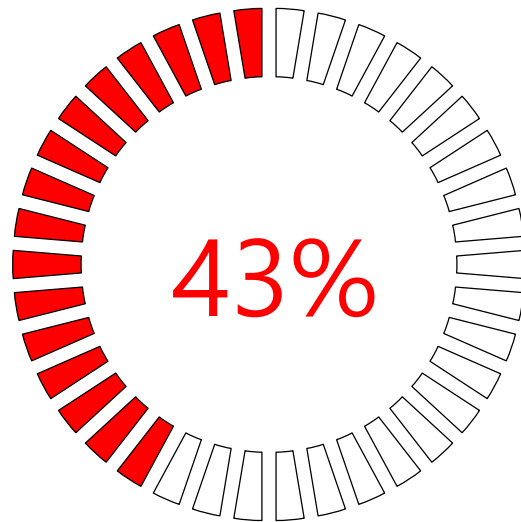
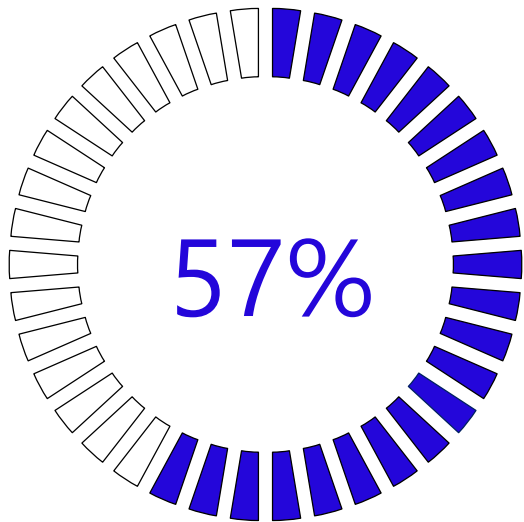


순위권 그룹 타이틀곡과 음원차트 주간순위 상관계수



	타이틀곡(가수)	상관계수
1	party(소녀시대)	-0.2
2	넌is원들(마마무)	-0.2
3	I Miss You(마마무)	-0.1
4	HOT PINK(EXID)	-0.1
5	lion heart(소녀시대)	0
6	Call me baby(엑소)	0.1
7	Love me right(엑소)	0.2
8	OOH-AHH하게 (트와이스)	0.3
9	Ah-Choo(러블리즈)	0.5

2차 결론



마무리

Q&A

감사합니다