Big Data Programming

AI 융합학부
20170403 최혜원

21.06.01

INDEX

1. 주제 요약

이 주제에 대한 설명

“정치 뉴스와 연예 뉴스의 상관관계 분석”

가설 : 어떤 연예계 사건이 터졌을 때, 정치적, 사회적 문제를 덮고자 하는 것일지도 모른다.

1. 주제 요약

구현 flow

If 특정 정치 키워드가 **이슈가 된 시기**를 파악

If 그 시기 전후로 어떤 **연예 관련 keyword 그래프가 급격히 상승**

Then 두 특정 키워드의 추이를 분석, 결과 도출

2. 수정1

구현 flow

If 특정 정치 키워드가 **이슈가 된 시기**를 파악

~~If 그 시기 전후로 어떤 **연예 관련 keyword 그래프가 급격히 상승**~~

Then ~~두 특정 키워드~~의 추이를 분석, 결과 도출

2. 수정1

구현 flow

If 특정 정치 키워드가 **이슈가 된 시기**를 파악

If 이슈가 된 시기 전후로 '연예' 카테고리의 volume 상승

Then 특정 정치 키워드와 연예계 뉴스 top 30간의 추이를 분석, 결과 도출

2. 수정1 이유

```
사드  의 최대 관심 날짜 ::  2016-07-01
startDate ::  2016-06-10
endDate ::  2016-07-22
NewsResult_20160610-20160722.xlsx
0                          김상근,단테,트웨인
1                                 이준관
2          김광보,브래드,김태훈,진,빌리,이창훈,케이틀린,칼
3                              고진,가라타니,아렌트
4              소무,전경욱,지족선사,전,노승,황진이,노장
                        ...
19995                             최강주
19996                             이강일
19997          정준영,여자친구,박나래,헨리,홍진영,에일리
19998          그리,니콜라스,레만,프랑수아즈,테일러,레픈
19999                          김해송,이난영
Name: 인물, Length: 20000, dtype: object
[('nan', 6743), ('이정재', 265), ('소녀시대', 247), ('연상호', 238), ('그리', 238), ('리암 니슨', 238), ('이재한', 233)
, ('오연주', 203), ('이종석', 193), ('한효주', 187), ('김', 169), ('샤이니', 163), ('여자친구', 162), ('맥아더', 154),
('구구단', 136), ('이범수', 133), ('트와이스', 133), ('진세연', 131), ('원더걸스', 128), ('마동석', 128), ('왕대륙', 12
0), ('홍상수', 117), ('에릭남', 117), ('김민희', 116), ('이', 107), ('송중기', 105), ('라미란', 105), ('김의성', 104),
('크나큰', 103), ('로미오', 102)]
```

- 이 결과만으로는 특정한 정치계 사건에 영향을 끼치는 특정한 연예 사건을 알 수 없다.
- 이 안에서 하나의 인물을 특정한다면 주관적 생각이 들어갈 수 있음이 우려되었다.

2. 수정1 이유

```
성완종 의 최대 관심 날짜 :: 2015-04-01
startDate :: 2015-03-18
endDate :: 2015-04-15
[('이태임', 816), ('유승옥', 697), ('유병재', 687), ('장동민', 672), ('수지', 662), ('엑소', 540), ('예원', 538), ('이문세', 514), ('이민호', 488), ('박진영', 482), ('김태우', 465), ('최현석', 425), ('강예원', 410), ('최여진', 405), ('그리', 401), ('김', 389), ('길건', 376), ('이병헌', 366), ('강균성', 364), ('미쓰에이', 344), ('최시원', 343), ('유재석', 335), ('박명수', 330), ('박유천', 328), ('김구라', 325), ('임지연', 323), ('유희열', 294), ('김강우', 291), ('안영미', 286)]
[{20150325: 1, 20150327: 66, 20150328: 175, 20150329: 139, 20150330: 184, 20150331: 80, 20150401: 53, 20150402: 50, 20150403: 21, 20150404: 5, 20150405: 5, 20150406: 15, 20150407: 1, 20150408: 2, 20150409: 3, 20150410: 1, 20150412: 3, 20150413: 11, 20150414: 1}, {20150325: 7, 20150326: 1, 20150327: 7, 20150328: 8, 20150329: 2, 20150330: 41, 20150331: 2, 20150401: 3, 20150402: 26, 20150403: 417, 20150404: 35, 20150405: 6, 20150406: 26, 20150407: 94, 20150408: 4, 20150409: 1, 20150410: 1, 20150411: 5, 20150412: 1, 20150413: 5, 20150414: 5}, {20150325: 7, 20150326: 1, 20150327: 4, 20150328: 32, 20150329: 102, 20150330: 26, 20150331: 6, 20150401: 2, 20150402: 4, 20150403: 2, 20150404: 15, 20150405: 5, 20150406: 5, 20150407: 4, 20150408: 321, 20150409: 20, 20150410: 3, 20150411: 109, 20150412: 6, 20150413: 3, 20150414: 10}, {20150325: 3, 20150326: 5, 20150327: 2, 20150328: 19, 20150329: 30, 20150330: 10, 20150331: 2, 20150401: 14, 20150402: 9, 20150403: 7, 20150404: 24, 20150405: 37, 20150406: 10, 20150407: 41, 20150408: 160, 20150409: 21, 20150410: 2, 20150411: 28, 20150412: 59, 20150413: 89, 20150414: 100}, {20150325: 43, 20150326: 20, 20150327: 18, 20150328: 7, 20150329: 9, 20150330: 110, 20150331: 13, 20150401: 15, 20150402: 26, 20150403: 242, 20150404: 19, 20150405: 20, 20150406: 58, 20150407: 16, 20150408: 15, 20150409: 4, 20150410: 1, 20150412: 1, 20150413: 17, 20150414: 8}, {20150325: 5, 20150326: 11, 20150327: 6, 20150328: 2, 20150329: 3, 20150330: 134, 20150331: 55, 20150401: 85, 20150403: 7, 20150404: 6, 20150405: 25, 20150406: 27, 20150407: 8, 20150408: 22, 20150409: 14, 20150410: 27, 20150411: 32, 20150412: 24, 20150413: 41, 20150414: 6}, {20150327: 50, 20150328: 144, 20150329: 64, 20150330: 126, 20150331: 53, 20150401: 30, 20150402: 12, 20150403: 12, 20150404: 10, 20150405: 10, 20150406: 12, 20150407: 3, 20150409: 2, 20150410: 1, 20150411: 1, 20150412: 6, 20150413: 2}, {20150325: 6, 20150326: 6, 20150327: 7, 20150330: 5, 20150331: 87, 20150401: 3, 20150402: 11, 20150403: 104, 20150404: 19, 20150406: 93, 20150407: 129, 20150408: 25, 20150409: 7, 20150410: 6, 20150411: 2, 20150412: 2, 20150414: 2}, {20150325: 54, 20150326: 28, 20150327: 25, 20150328: 7, 20150329: 7, 20150330: 52, 20150331: 11, 20150401: 14, 20150402: 18, 20150403: 143, 20150404: 40, 20150405: 9, 20150406: 22, 20150407: 12, 20150408: 26, 20150409: 15, 20150410: 2, 20150413: 2, 20150414: 1}, {20150326: 2, 20150327: 9, 20150328: 2, 20150329: 24, 20150330: 29, 20150331: 7, 20150401: 5, 20150403: 2, 20150405: 18, 20150406: 6
```

2. 수정2

구현 flow

- 트위터의 파이썬 라이브러리인 **tweepy를 활용**하여 과거 및 현재 인기 트위터 데이터를 가져와서 가공하였다.
- Tweet volume 이 일정 수 이상 큰 경우의 단어를 모으기 위해 volume 값들을 출력하였다.

- 날짜는 cursor 메서드의 since값을 지정하여 원하는 날짜의 data를 가져올 수 있다.

- Tweeter 데이터를 가져와서 확인해보았더니 **비속어, 신조어** 등 내용을 정확히 파악할 수 없는 데이터들이 많았다.
- Volume값에 **null 값**이 많았다.

3. 구현 방법

구현 flow - 정치

If 특정 정치 키워드가 이슈가 된 시기를 파악

1) 키워드를 정한다. (ex. 국정원 해킹, 교과서 국정화 추진 논란, LH)
2) 정한 키워드를 엑셀 파일에 입력한다.
3) Google trend 를 이용하여 2010-01-01 ~ 2021-01-01 사이의 엑셀 파일에 입력된 단어 중 가장 많이 검색된 날짜를 출력한다.
4) 해당 날짜 전, 후로 2주를 기간으로 두고 연예계 뉴스를 파악한다.

3. 구현 방법

구현 flow - 정치

```
for idx, key in enumerate(keywords):
    # period = str(data[idx][0]) + ' ' + str(data[idx][1])
    pytrends.build_payload(key, cat = 0, timeframe ='2010-01-01 2021-01-01', geo = 'KR', gprop = '')
    getDataInfo = pytrends.interest_over_time()
```

구글 트렌드에서 키워드에 대한 데이터 정보를 가져옴

```
# 키워드의 최고 vol날짜를 찾는 것
for i in key:
    maxVal = max(getDataInfoCsv[i])
    for idx, valVol in enumerate(getDataInfoCsv[i]):
        if(maxVal == valVol):
            maxVolDate.append(getDataInfoCsv.iloc[idx]["date"])
```

가져온 데이터의 최대 검색량이 도출된 날짜를 가져옴

```
for idx, key in enumerate(keywords):
    print(key[0], ' 의 최대 관심 날짜 :: ' , maxVolDate[idx])
    year = int(maxVolDate[idx][:4])
    month = int(maxVolDate[idx][5:7])
    day = int(maxVolDate[idx][8:])

    time = datetime(year,month,day)
    startDate = str(time + timedelta(weeks =-3))[:10].strip()
    endDate = str(time + timedelta(weeks = 3))[:10].strip()
    print("startDate :: ", startDate )
    print("endDate :: ", endDate )
```

해당 날짜의 2주 전, 후의 날짜를 계산

3. 구현 방법

구현 flow - 연예

5) 해당 시기 동안 모든 연예 관련 **뉴스 데이터들을 수집**한다.
6) 각 뉴스에 해당하는 인물 column을 불러와 빈도수에 맞게 수집, 정렬한다.
7) 정렬된 데이터들 중 **top 30을 추출**한다.
8) 그 데이터가 호출된 시기, 즉 top 30 각각의 인물 키워드가 나온 **시기별로 정리**하여 그래프로 시각화한다.

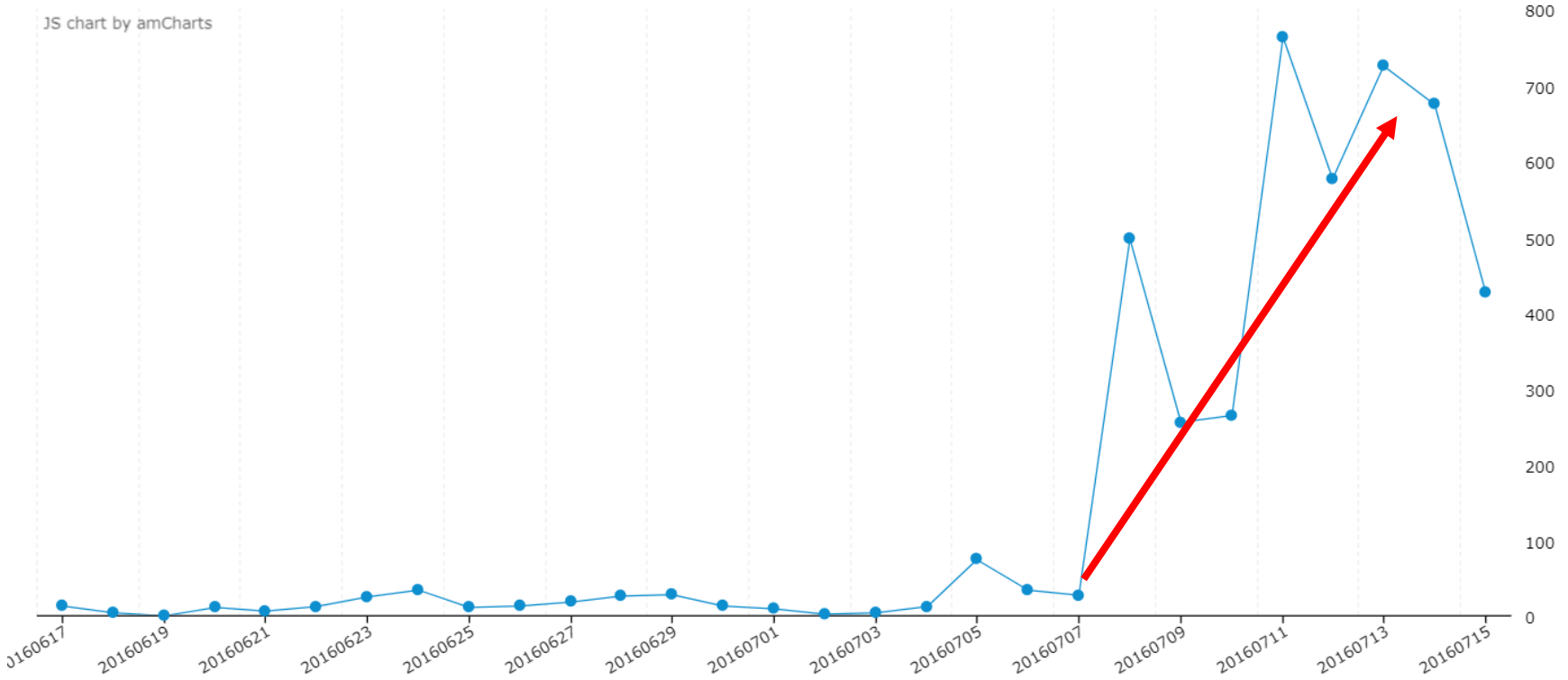
4. 결과

정치	연예계	시작	종료
사드			
최순실			
국정원 댓글			

```
사드  의 최대 관심 날짜 ::  2016-07-01
startDate ::  2016-06-17
endDate ::  2016-07-15
최순실  의 최대 관심 날짜 ::  2016-10-01
startDate ::  2016-09-17
endDate ::  2016-10-15
국정원 댓글  의 최대 관심 날짜 ::  2013-08-01
startDate ::  2013-07-18
endDate ::  2013-08-15
```

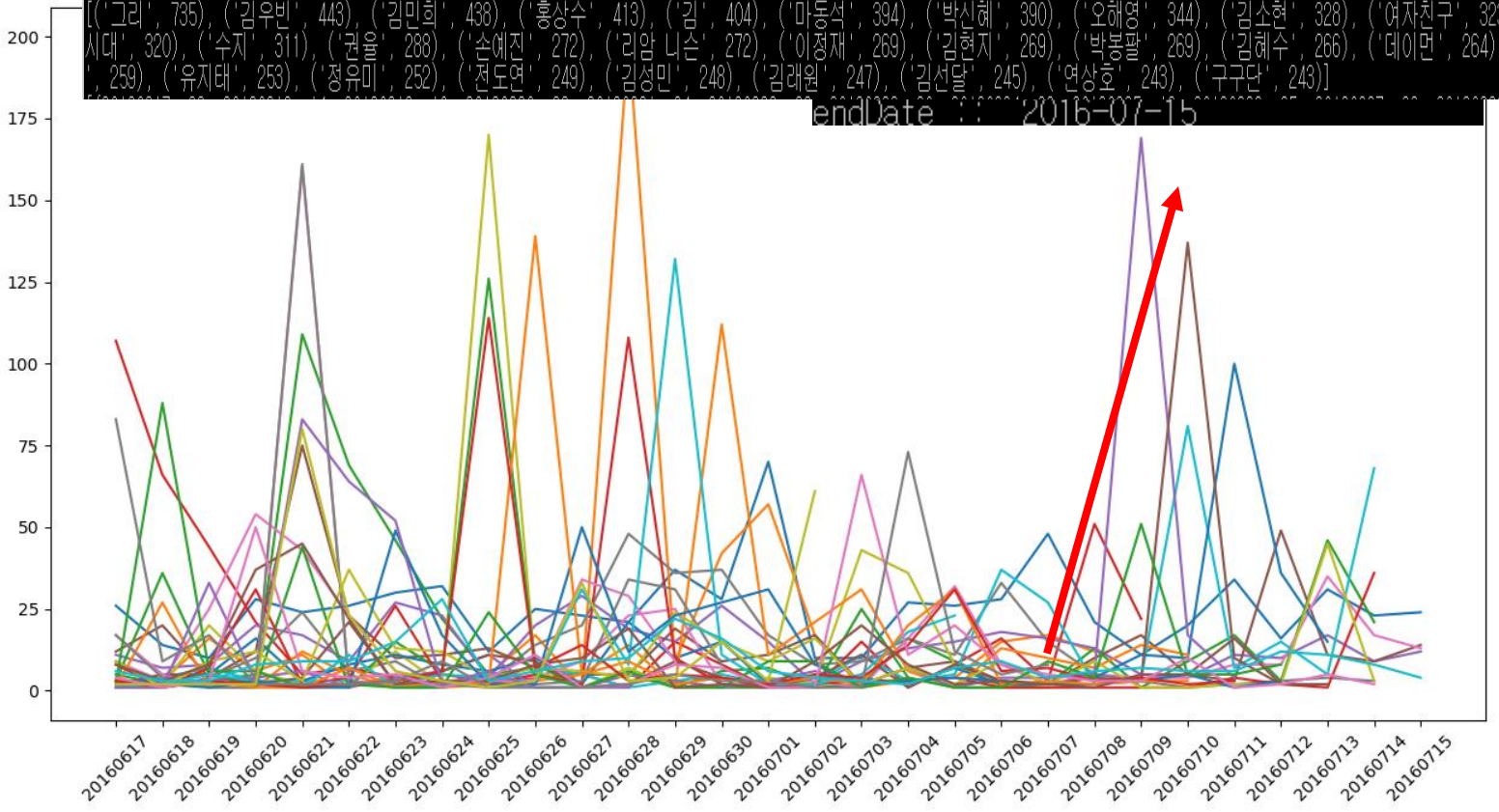
4. 결과

' 사드'라는 키워드로 동일 기간에 분석하였을 때 뉴스 수



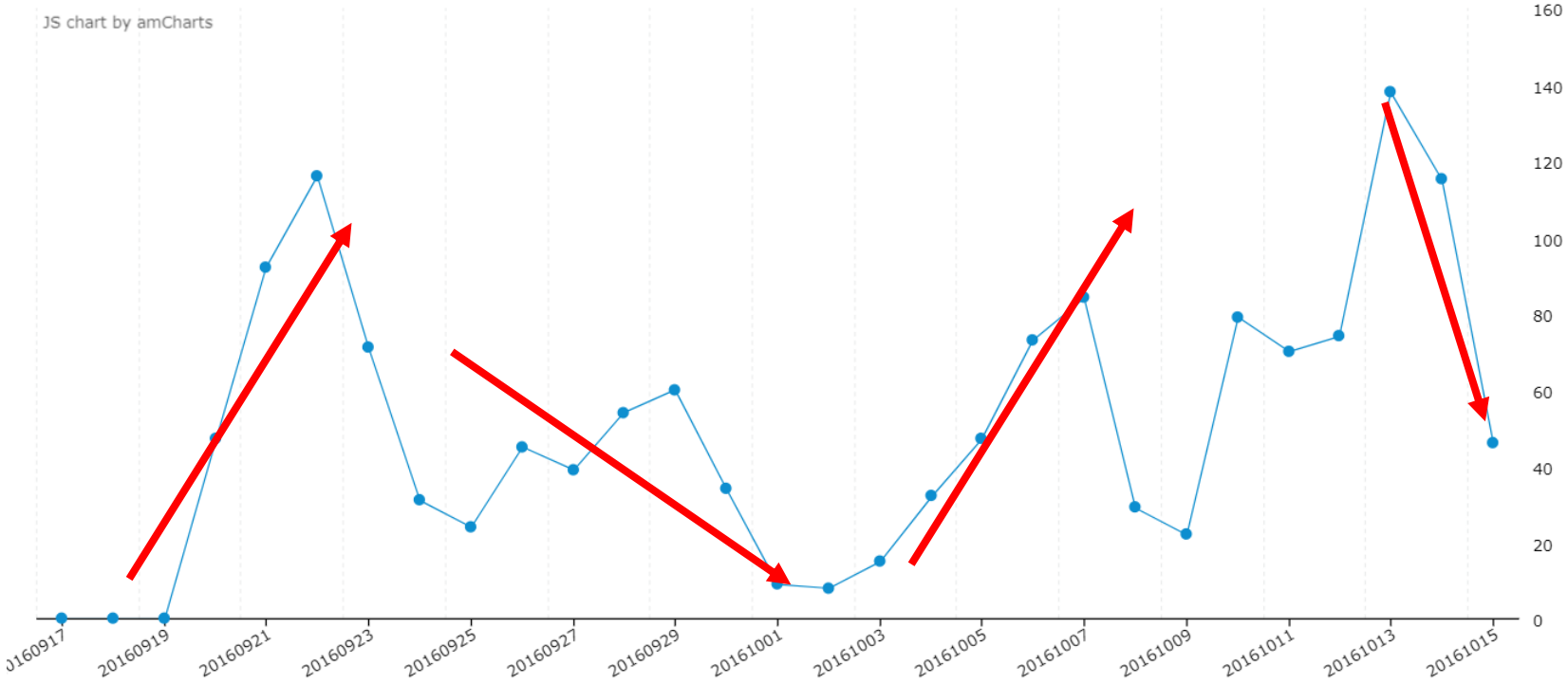
4. 결과

```
사드 의 최대 관심 날짜 :: 2016-07-01
startDate :: 2016-06-17
endDate :: 2016-07-15
[('그리', 735), ('김우빈', 443), ('김민희', 438), ('홍상수', 413), ('김', 404), ('마동석', 394), ('박신혜', 390), ('오해영', 344), ('김소현', 328), ('여자친구', 323), ('소녀
시대', 320), ('수지', 311), ('권율', 288), ('손예진', 272), ('리암 니슨', 272), ('이정재', 269), ('김현지', 269), ('박봉팔', 269), ('김혜수', 266), ('데이먼', 264), ('조진웅
', 259), ('유지태', 253), ('정유미', 252), ('전도연', 249), ('김성민', 248), ('김래원', 247), ('김선달', 245), ('연상호', 243), ('구구단', 243)]
```



4. 결과

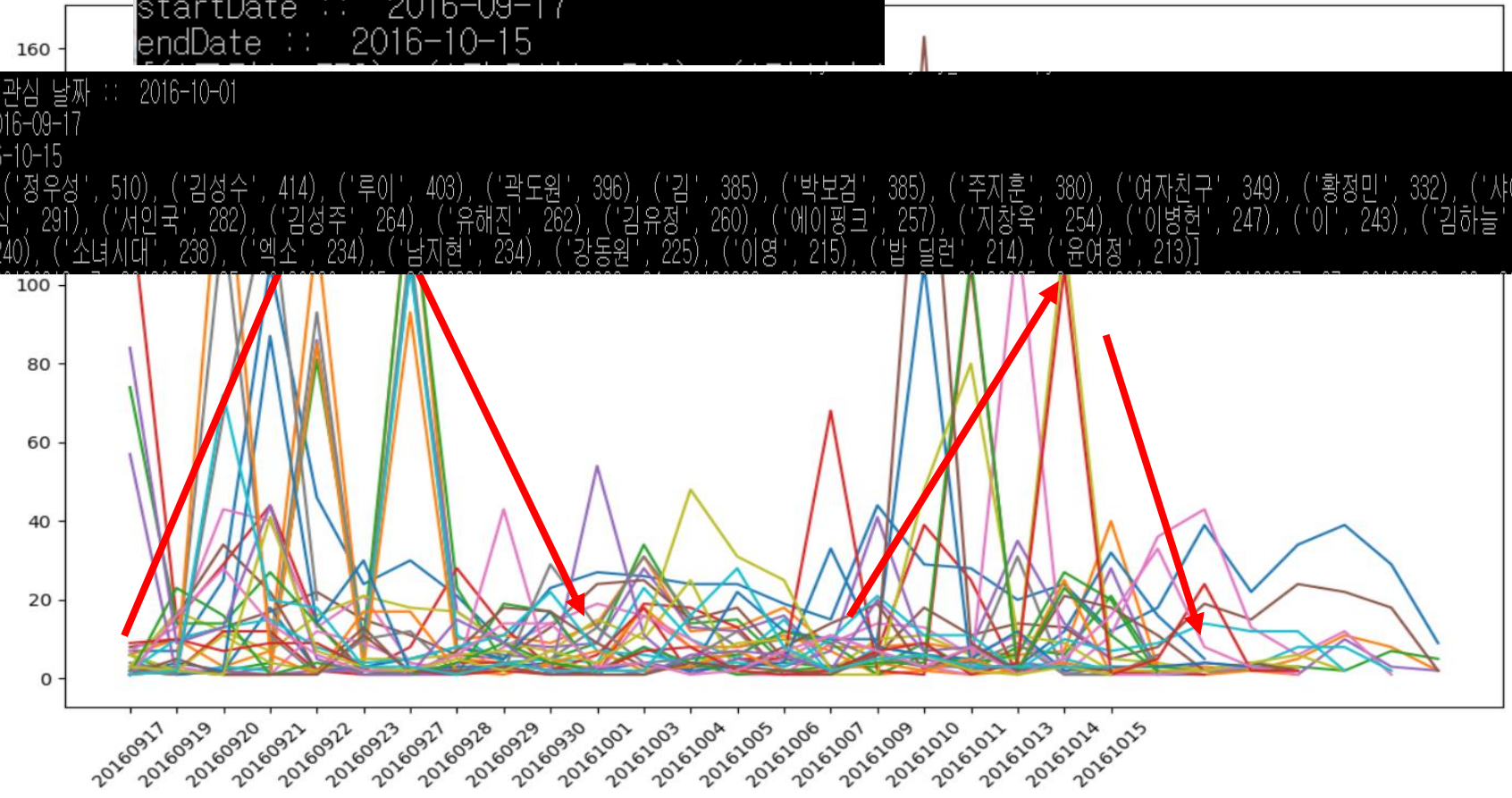
' 최순실' 이라는 키워드로 동일 기간에 분석하였을 때 뉴스 수



4. 결과

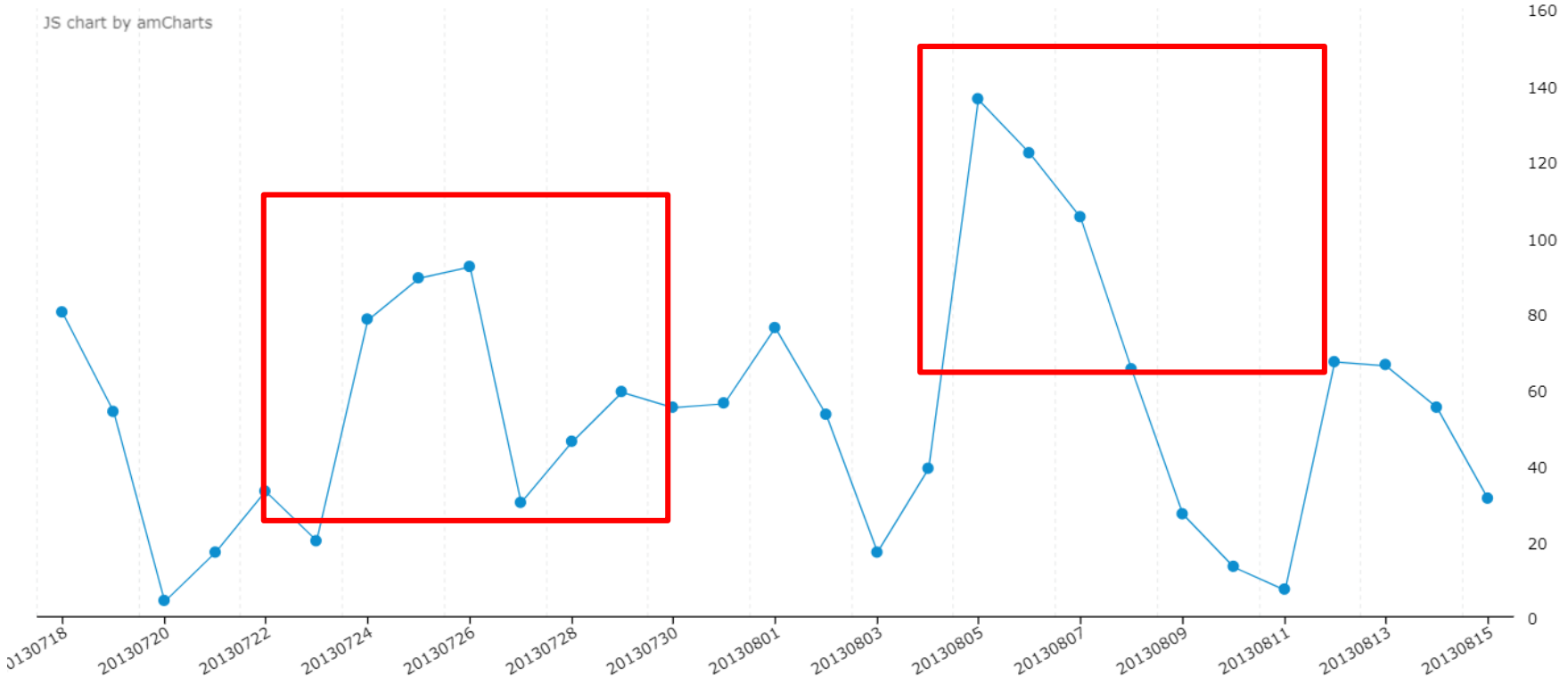
```
최순실  의 최대 관심 날짜 ::  2016-10-01
startDate ::  2016-09-17
endDate ::  2016-10-15
```

```
최순실  의 최대 관심 날짜 ::  2016-10-01
startDate ::  2016-09-17
endDate ::  2016-10-15
[('그리', 779), ('정우성', 510), ('김성수', 414), ('루이', 403), ('곽도원', 396), ('김', 385), ('박보검', 385), ('주지훈', 380), ('여자친구', 349), ('황정민', 332), ('샤이니', 306), ('정만식', 291), ('서인국', 282), ('김성주', 264), ('유해진', 262), ('김유정', 260), ('에이핑크', 257), ('지창욱', 254), ('이병헌', 247), ('이', 243), ('김하늘', 242), ('송윤아', 240), ('소녀시대', 238), ('엑소', 234), ('남지현', 234), ('강동원', 225), ('이영', 215), ('밥 딜런', 214), ('윤여정', 213)]
```



4. 결과

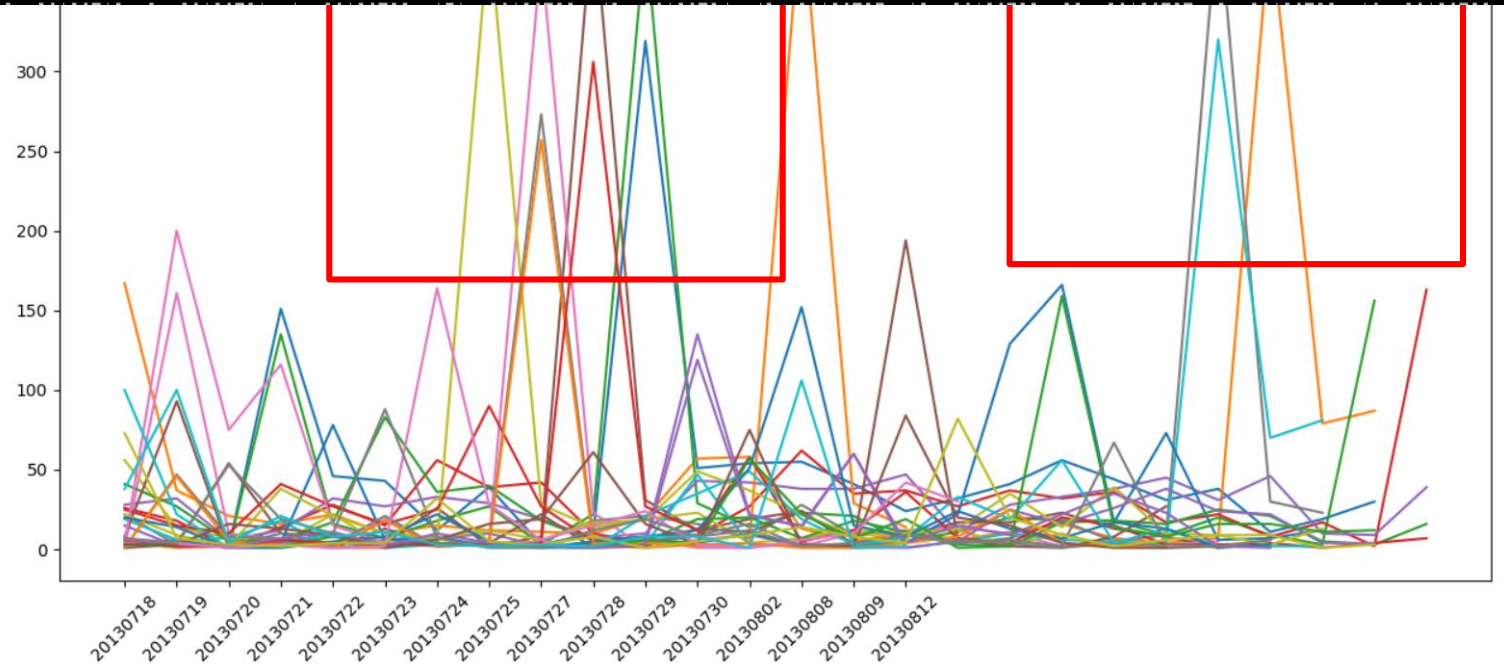
' 국정원 댓글'이라는 키워드로 동일 기간에 분석하였을 때 뉴스 수



4. 결과

```
국정원 댓글  의 최대 관심 날짜 ::  2013-08-01
startDate ::  2013-07-18
endDate ::  2013-08-15
```

```
[('봉준호', 1160), ('이병헌', 1095), ('송강호', 978), ('이종석', 784), ('그리', 733), ('틸다 스윈튼', 700), ('고아성', 682), ('신동엽', 666), ('크리스 에반스', 647), ('이민정', 624), ('장혁', 575), ('이범수', 486), ('김', 485), ('클라라', 477), ('소지섭', 457), ('수애', 441), ('김종학', 438), ('존 허트', 432), ('하정우', 425), ('이효리', 411), ('이보영', 406), ('제이미 벨', 396), ('손현주', 384), ('신영균', 378), ('공효진', 377), ('에프엑스', 373), ('문채원', 347), ('아이유', 346), ('걸스데이', 323)]
```

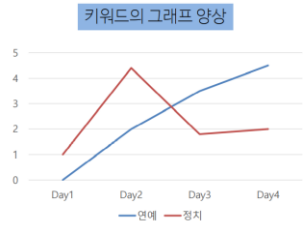


5. 결과분석

결과들을 분석한 결과

1. 특정 정치적 사건 뉴스 양이 많아짐에 따라 연예계 뉴스 양도 많아짐을 알 수 있었다.

2. 특정 정치 키워드의 최대 관심 날짜에는 연예계 뉴스 척도가 낮아졌다.

3. 

이와 같은 모양은 아니었다.

1. Top 30 안에 특정 정치적 사건 키워드가 들어있지 않음에도 이슈 양상이 비슷하기에, 정치 사건이 연예 사건에 **어느 정도 '영향을 끼친다'**는 것을 알 수 있다.

2. 그 당시에는 연예 뉴스보다는 정치 뉴스에 더 관심이 많아짐을 알 수 있다.

6. 결론

“정치 뉴스와 연예 뉴스의 상관관계 분석”

-〉 어느 정도 영향이 있다.

QnA

Thank you