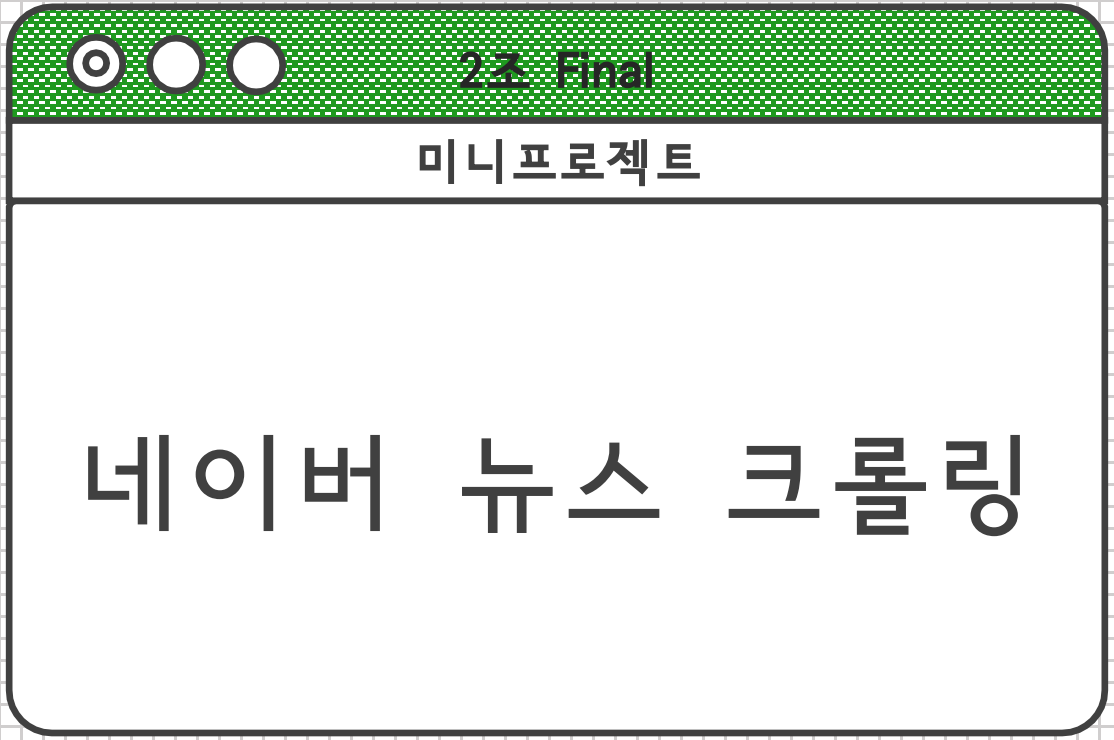


최민성 최근호 안지영

개요

네이버 뉴스를 2019.1.1~2019.12.31 크롤링 한 후

Wordcloud를 통해 그 달의 이슈를 파악하고자 함

목차

◎○○ 크롤링

Beautiful soup 패키지 이용

In [5]:
```
news_headline=[]
for n in range(1,366):
    date_url = url.format(date[n])
    html = urlopen(date_url)
    soup = BeautifulSoup(html, 'html.parser')
    titles_html = soup.select('.ranking_section > ol > li > dl > dt > a')

    for a in range(len(titles_html)):
        news_headline.append(titles_html[a].text)

    n = n+1
```

In [6]:
```
print(news_headline)
```

```
["지만원 '거짓 판명' 5·18 北 개입설 또 반복…비난 쇄도", '한국당 "손혜원 국조 받아라" …민주 "의혹 있는 모든의원 다…', '"정책 나올때마다 시장왜곡 우려" "주52시간도 또하나의 규…', '"朴 전대통령 건강 좋지 않지만… 위독설·체중39㎏은 사실…', '文대통령 "예타 제도 유지…균형발전 위해 개선 필요 있어"', '[벼랑끝 공인중개사②]온·오프라인서 동반 위협…거래절벽…', '작년 세금 계획보다 25.4조 더 걷혔다…역대 최대 초과 세수', '초과 세수 25.4조 원 '사상 최대' …"반도체·부동산 호황 덕…', '韓 외환보유액 4055억 달러 사상 최대…석달 연속 증가세(종…', '"내 택배 어디쯤 왔나"…₩'택배 조회₩' 종일 실검에 왜?', '이번엔 뚝섬 장어집 "골목식당 제작진이 사기꾼 만들어" … …', "전국 대부분 지역 '한파주의보'…서울 체감온도 영하 14도", '정부와 정치권에 경종 울린 두 의사의 죽음 [어떻게 생각하…', '서울 국공립 중고교 교사 임용에 836명 합격…남성은 23%', '"한 명이라도 더 살려야"…끝까지 현장 지킨 ₩'응급의료 버팀…', '[날씨] 출근길 동장군 맹위…체감온도 영하 15도', "[날씨] 전국 대부분 '한파주의보'...주말에도 강추위 계속", '설 연휴 인천공항 하루평균 이용객 20만명 넘어…역대 최다', '"너도 스테로이드 했지?"…피트니스계에 번지는 ₩'약투₩'', '[날씨] 주말·휴일 내내 한겨울 추위…건조특보 확대', "베네수엘라, 물고 물린 '쩐의 전쟁'", "2월 미중 정상회담 무산 배경은?…'북미회담과 분리 의도'", '[특파원리포트] 대단한 중국인들…방콕 호텔 투숙하며 '원정…', '日언론 "문 대통령, ₩'징용공 배상은 日기업 문제₩' 입장"', "크로아티아서 2억원 넘는 '장어 밀반출' 한국인 2명 체포", "LG發 유료방송 '지각변동'…KT·SKT도 속속 케이블TV 인…", '[이진욱의 전자수첩] 한국은
```

◎○○ 전처리

크롤링 raw데이터는 list

List -> dataframe
(csv 저장을 위해)

In [7]:
```
data = pd.DataFrame(news_headline)
#list는 csv로 저장이 안되어서 dataframe으로 변환
```

In [8]:
```
data=data.reset_index(drop=True)
data
```

Out [8]:

	0
0	지만원 '거짓 판명' 5·18 北 개입설 또 반복…비난 쇄도
1	한국당 "손혜원 국조 받아라"…민주 "의혹 있는 모든의원 다…
2	"정책 나올때마다 시장왜곡 우려" "주52시간도 또하나의 규…
3	"朴 전대통령 건강 좋지 않지만… 위독설·체중39kg은 사실 …
4	文대통령 "예타 제도 유지…균형발전 위해 개선 필요 있어"
...	...
10945	잇몸에서 속눈썹 털이 자라는 여성
10946	[자막뉴스] 면역력에 '독'…이 2가지를 조심하세요!
10947	[속보] 신종 코로나 확진자 1명 추가…우한 교민 중 1명
10948	'갤럭시S20 울트라' 실물 이미지 유출…"삼성 스마트 냉장고…
10949	NS홈쇼핑, 8·9일 KF94 마스크 방송 편성

10950 rows × 1 columns

◎○○ 전처리

csv -> txt 전환 후 load

List -> string 변환

```
In [28]: news1

Out[28]: '["文대통령 당선 일등공신 ₩'광흥창팀₩', 시련의 겨울", ₩' "탈북민 10여명 베트남 당
국에 체포, 中 추방…韓대사관 도…₩', ₩' "韓 대사관, 기다리라더니…" 탈북민 10명
베트남서 체포돼 …₩', ₩'여야 '민식이법 눈물'에 화들짝… 한국당 원인 제공 목소리
₩', ₩'샘 해밍턴의 금강산 방문…빗장 연 北, 의도는₩', "아버지 채무 때문에 ₩'상속포
기₩' · · ·내 딸이 그 빚 물려받…", ₩'887회 로또당첨번호 조회 결과 서울 1등 3명 최
다…전체 38%₩', ₩'8명이 각각 25억원씩…제887회 로또당첨번호는?₩', ₩' "삼성전자를
지금이라도 매수할까요?" ₩', ₩'중국인 입맛 사로잡은 한국 라면…수입라면 1위₩',
₩' '성남어린이집 성추행 의혹' 일파만파…가해자 처벌 국민 청…₩', ₩'강남 외제차에
연봉1억도 "난 가난"… 대한민국 진짜 가난은…₩', ₩'서울 도심 5등급차량 제한 개시 7
시간만에 205대 단속₩', ₩'"성남시 어린이집 성폭행 아동 부모, 국가대표 박탈해야"
靑…₩', ₩'허벅지 만지며 추행 뒤 "순수한 사랑" · · ·그녀는 대자보 …₩', ₩'신혼집 텐
트 짊어지고 세계로 · · ·4년째 신혼여행 중인 이 …₩', ₩'[사소한 발견] 여왕은 왜 드
레스를 벗어던졌나 · · ·레깅스 …₩', ₩'짐 로저스 로저스홀딩스 회장 "내년 세계경제
큰 위기 온다…₩', ₩'화순 주차장 사건의 진짜 최종결말₩', "₩'해도 해도 너무하네₩' 부
산 모 여고 경비원 채용공고 분노 …", ₩'입으로 소변 800㎖ 받아냈다, 비행기서 노인
살린 中의사₩', ₩' '살인미수→가정폭력' 혐의 낮춰 남편 풀려난 뒤 살해된 아…₩',
₩'총 맞은 아기상어 인형…3살 아기 목숨 구했다₩', ₩'日 교도통신 "아베 총리, UN서 연
설하려다 거부당해"₩', ₩'인도서 또 잔혹 성폭행 · 살인…수천명 "범인 넘겨라" 항의 …
₩', ₩'기대 이상 &amp;apos;갤폴드&amp;apos;, 출시국 2배↑…새 모델 2…₩', ₩'₩₩₩'완판₩₩₩' 갤
```

◎○○ 자연어처리 패키지

nltk

Natural Language
Analyses with NLTK

konlpy

Konlpy는 오직 “한국어”를 위한 패키지이나, 설치가 되지 않았음

◎○○ 전처리

특수문자 제거와,
단어분리를 위해 tokenize

특문제거 X
tokenize 제대로 안됨

-> Sent_tokenize 실패

sent_tokenize는 실패

In [15]:
```
from nltk import sent_tokenize
text=sent_tokenize(news)
```

In [29]:
```
text
```

Out [29]:
```
['["文대통령 당선 일등공신 ₩'광흥창팀₩', 시련의 겨울", ₩' "탈북민 10여명 베트남 당국에 체포, 中 추방…韓대사관 도…₩', ₩' "韓 대사관, 기다리라더니…" 탈북민 10명 베트남서 체포돼 …₩', ₩'여야 '민식이법 눈물' 에 화들짝… 한국당 원인 제공 목소리₩', ₩'샘 해밍턴의 금강산 방문…빗장 연 北, 의도는₩', "아버지 채무 때문에 ₩'상속포기₩' · · ·내 딸이 그 빚 물려받…", ₩'887회 로또당첨번호 조회 결과 서울 1등 3명 최다…전체 38%₩', ₩'8명이 각각 25억원씩…제887회 로또당첨번호는?',
 '₩', ₩' "삼성전자를 지금이라도 매수할까요?" ₩', ₩'중국인 입맛 사로잡은 한국 라면…수입라면 1위₩', ₩' '성남어린이집 성추행 의혹' 일파만파…가해자 처벌 국민 청…₩', ₩'강남 외제차에 연봉1억도 "난 가난"… 대한민국 진짜 가난은…₩', ₩'서울 도심 5등급차량 제한 개시 7시간만에 205대 단속₩', ₩'"성남시 어린이집 성폭행 아동 부모, 국가대표 박탈해야" 靑…₩', ₩'허벅지 만지며 추행 뒤 "순수한 사랑" · · ·그녀는 대자보…₩', ₩'신혼집 텐트 짊어지고 세계로 · · ·4년째 신혼여행 중인 이 …₩', ₩'[사소한 발견] 여왕은 왜 드레스를 벗어던졌나 · · ·레깅스 …₩', ₩'짐 로저스 로저스홀딩스 회장 "내년 세계경제 큰 위기 온다…₩', ₩'화순 주차장 사건의 진짜 최종결말₩', "₩'해도 해도 너무하네₩' 부산 모 여고 경비원 채용공고 분노 …", ₩'입으로 소변 800㎖ 받아냈다, 비행기서 노인 살린 中의사₩', ₩' '살인미수→가정폭력' 혐의 낮춰 남편 풀려난 뒤 살해된 아…₩', ₩'총 맞은 아기상어 인형…3살 아기 목숨 구했다₩', ₩'日 교도통신 "아베 총리, UN서 연설하려다 거부당해"₩', ₩'인도서 또 잔혹 성폭행 · 살인…수천명 "범인 넘겨라" 항의 …₩', ₩'기대 이상 &apos;갤폴드&apos;, 출시국 2배↑…새 모델 2…₩',
```

◎○○ 전처리

Regex를 이용한
토큰화 성공!

한글을 제외한 나머지 제거

```
from nltk.tokenize import RegexpTokenizer
re_capt = RegexpTokenizer('[가-힣]ww+')
```

```
In [20]: news2

Out[20]: ['대통령',
          '당선',
          '일등공신',
          '광흥창팀',
          '시련의',
          '겨울',
          '탈북민',
          '여명',
          '베트남',
          '당국에',
          '체포',
          '추방',
          '대사관',
          '대사관',
          '기다리라더니',
          '탈북민',
          '베트남서',
          '체포돼',
          '여야',
```

전처리

불용어 사전 처리

stop_words = "한국 속보 종합 단독 논란 날씨 만원 대통령"

figure 비교

fig1 <크롤링한 raw data>

fig2 < regex를 이용한 tokenize>

figure 비교

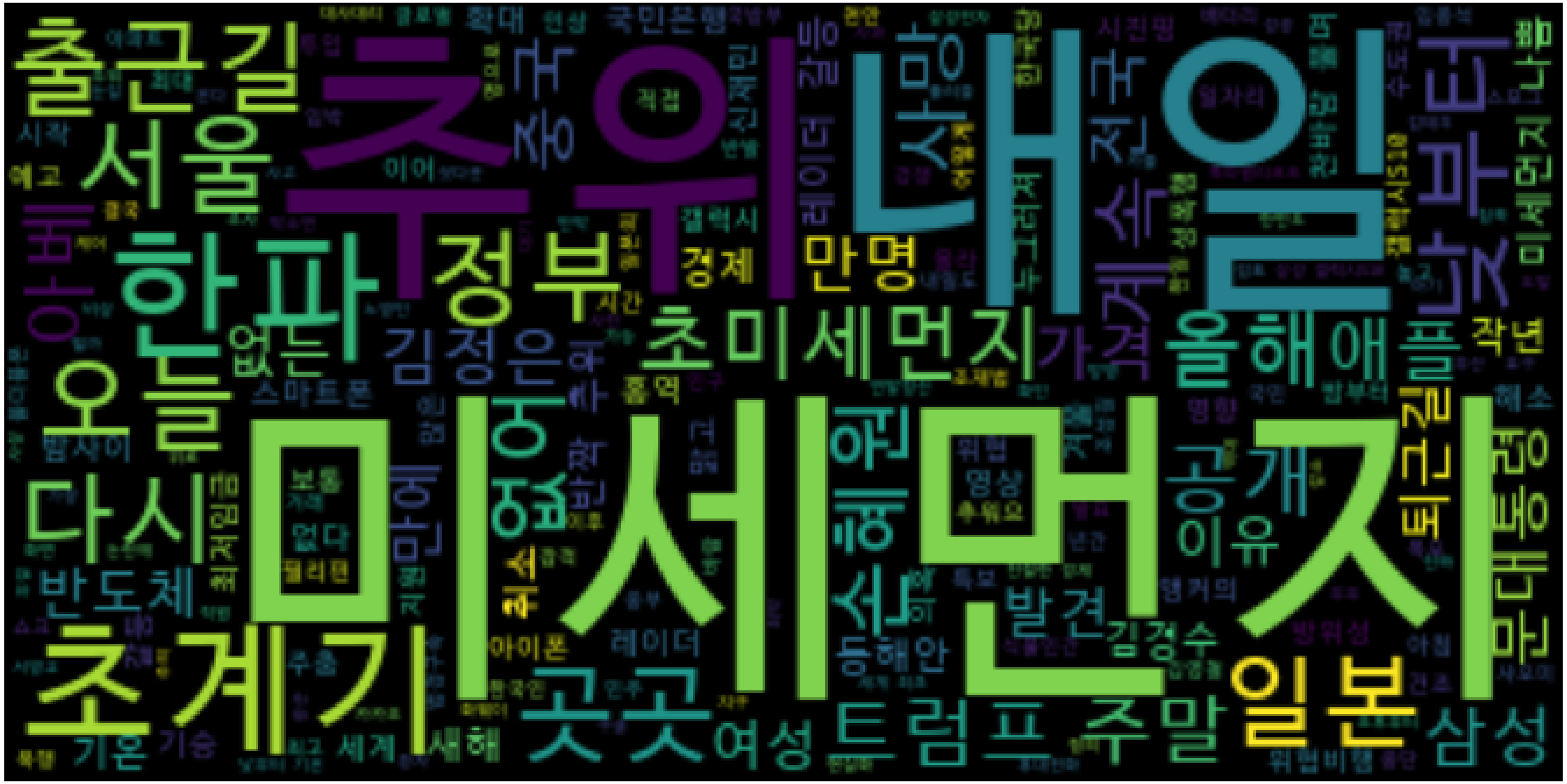
fig2 <regex를 이용한 tokenize>

fig3 <stop words 적용>


BIT
REWIND
2019

결과

19년 1월의 이슈



미세먼지
내일
추위
출근길
서울
한파
오늘
다시
초계기
정부
초미세먼지
올해
애플
곳곳
트럼프
주말
일본
삼성
여성
김정은
반도체
만명
경제
발견
이유
가격
작년
스마트폰
확대
국민은행
시진핑
레이더
김경수
등해안
아이폰
홍역

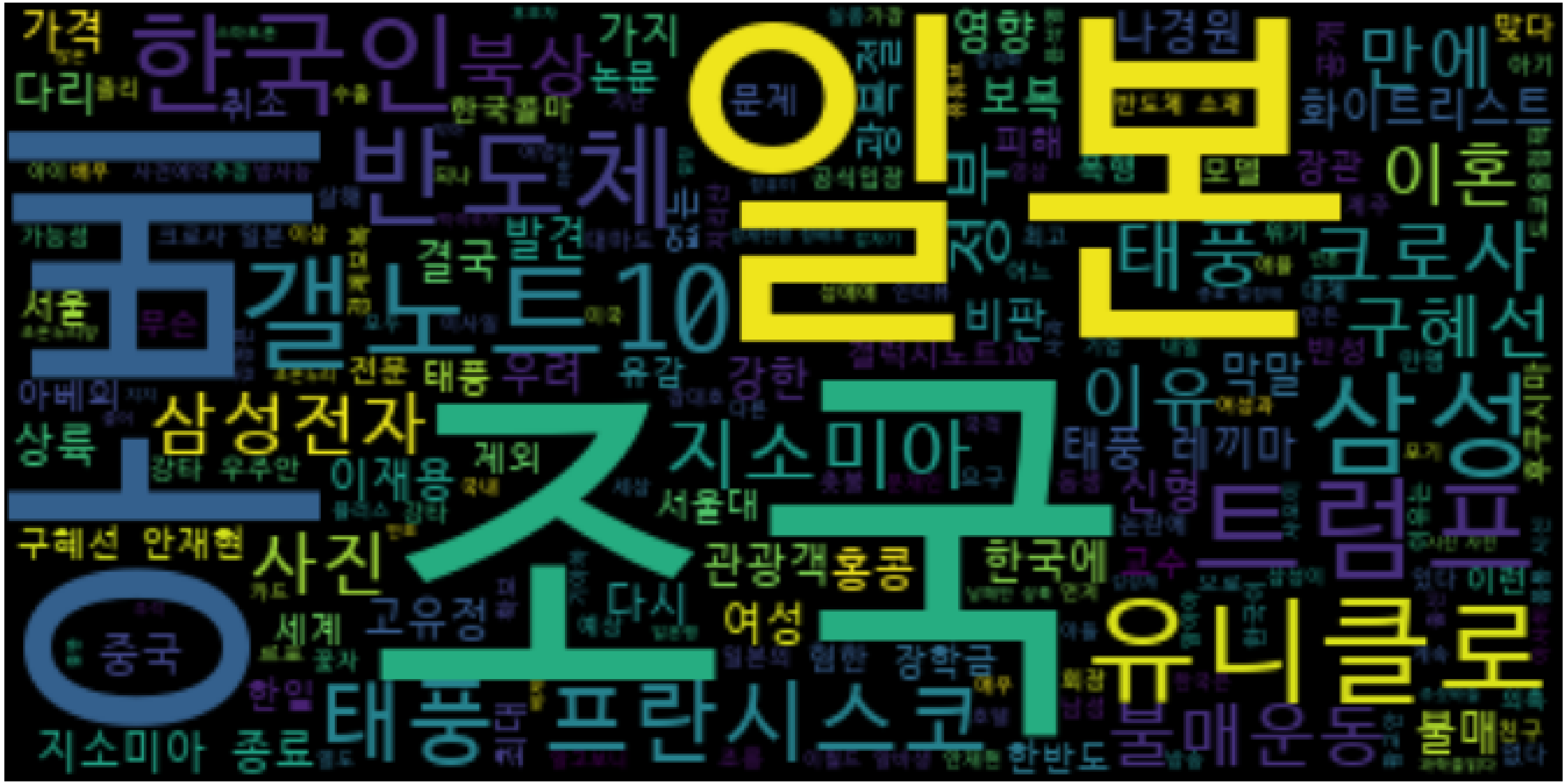
결과

19년 3월의 이슈


전국
승리
미세먼지
내일
중국
삼성
주말
정부
애플
비상저감조치
포근
먼지
출근길
갤럭시S10
박영선
제주
의혹
충북
베트남
카풀
밤사이
마스크
계속
후폭풍
나경원
대한항공
꽃샘추위
서울
일부
김정은
경찰
만에
비상
서쪽
사망
동창리
오늘
최대
요란한
곳곳
시간
이희진
구속
사고
아침
추락
내륙
정준영
갤S10
가장
대기질
쌀쌀
미세먼지
나쁨
퇴근길

결과

19년 8월의 이슈



일본
조국
홍콩
한국인
반도체
갤노트10
트럼프
삼성
유니클로
태풍
프란시스코
삼성전자
지소미아
불매운동
크로사
구혜선
화이트리스트
이혼
이유
만에
나경원
북상
가격
다리
가지
논문
영향
보복
중국
사진
관광객
고유정
서울대
이재용
한국에
강한
우려
발견
결국
비판
막말
태풍 레끼마
여성
장학금
다시
세계
한일
지소미아 종료
구혜선 안재현
상륙
서울
아베의
전문
태풍
한반도
물매

결과

19년 12월의 이슈

19년 전체의 핫 이슈 (19.1.1~ 19.12.31)


Thank you