윈도우 환경에서의 웹크롤링을 통한 악성코드 검사

바이러스토탈 API DB연동과 광고성 웹크롤링

SAMZO

1771064 김연희 1771070 안해인
1771071 유예린 1771077 전재원

SAMZO

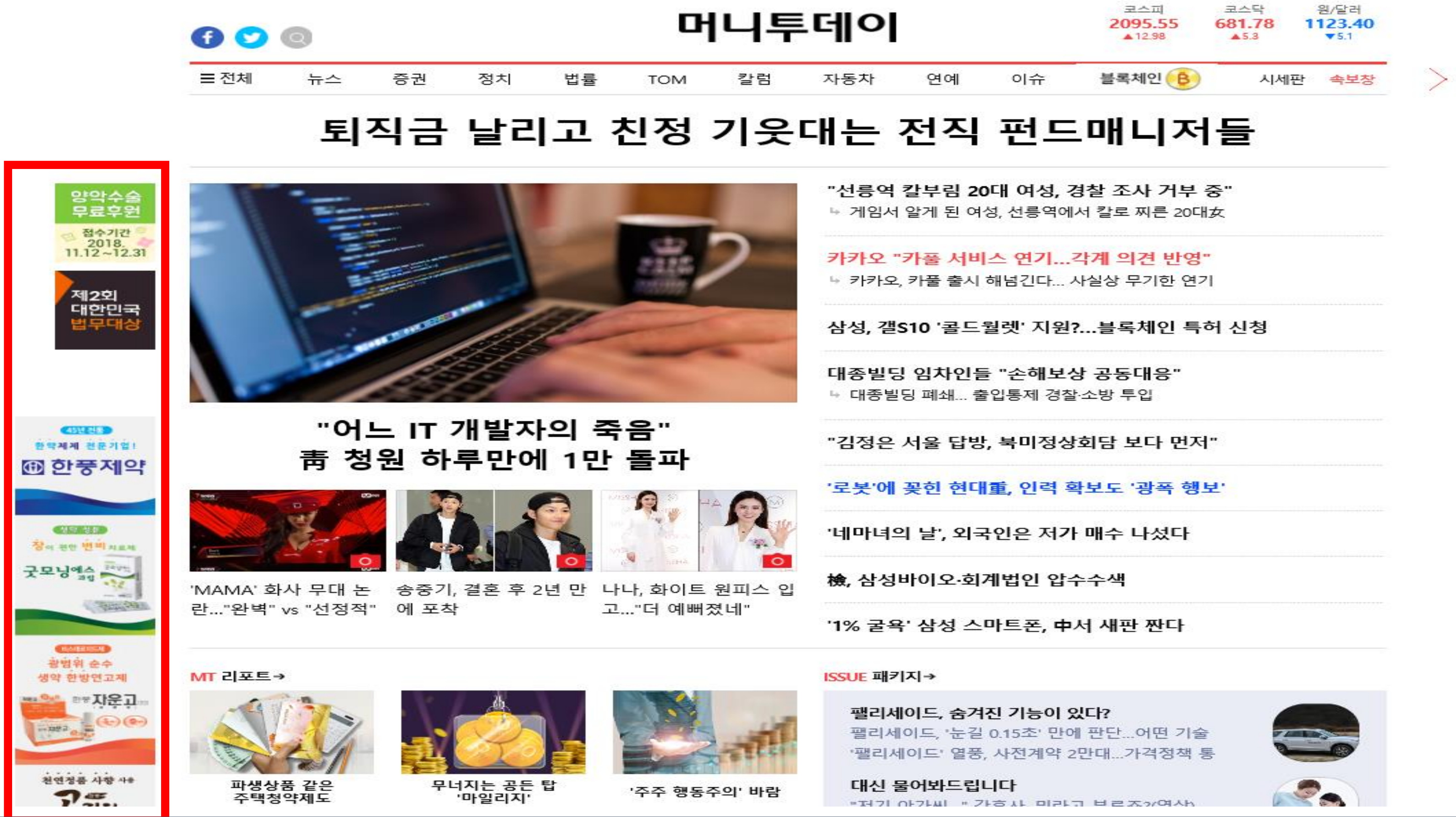
목차

SAMZO

1. 프로젝트 소개

SAMZO

[구현목적]



머니투데이

코스피 2095.55 ▲12.98 | 코스닥 681.78 ▲5.3 | 원/달러 1123.40 ▼5.1

≡ 전체 뉴스 증권 정치 법률 TOM 칼럼 자동차 연예 이슈 블록체인 시세판 속보창

퇴직금 날리고 친정 기웃대는 전직 펀드매니저들

"어느 IT 개발자의 죽음"
靑 청원 하루만에 1만 돌파

'MAMA' 화사 무대 논란..."완벽" vs "선정적"

송중기, 결혼 후 2년 만에 포착

나나, 화이트 원피스 입고..."더 예뻐졌네"

"선릉역 칼부림 20대 여성, 경찰 조사 거부 중"
ㄴ 게임서 알게 된 여성, 선릉역에서 칼로 찌른 20대女

카카오 "카풀 서비스 연기...각계 의견 반영"
ㄴ 카카오, 카풀 출시 해넘긴다... 사실상 무기한 연기

삼성, 갤S10 '콜드월렛' 지원?...블록체인 특허 신청

대종빌딩 임차인들 "손해보상 공동대응"
ㄴ 대종빌딩 폐쇄... 출입통제 경찰·소방 투입

"김정은 서울 답방, 북미정상회담 보다 먼저"

'로봇'에 꽂힌 현대重, 인력 확보도 '광폭 행보'

'네마녀의 날', 외국인은 저가 매수 나섰다

檢, 삼성바이오·회계법인 압수수색

'1% 굴욕' 삼성 스마트폰, 中서 새판 짠다

MT 리포트→

파생상품 같은 주택청약제도

무너지는 곤든 탑 '마일리지'

'주주 행동주의' 바람

ISSUE 패키지→

팰리세이드, 숨겨진 기능이 있다?
팰리세이드, '눈길 0.15초' 만에 판단...어떤 기술
'팰리세이드' 열풍, 사전계약 2만대...가격정책 통

대신 물어봐드립니다

양악수술 무료후원
접수기간 2018. 11.12~12.31

제2회 대한민국 법무대상

45년 전통 한약제제 전문기업! 한풍제약

장이 편한 변비 치료제 굿모닝에스

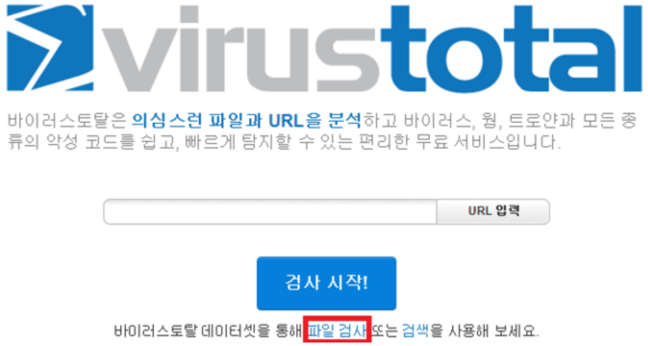
광범위 순수 생약 한방연고제 자운고

천연정품 사향

SAMZO

[시나리오]

1. 임의의 웹사이트 크롤링
→ 광고성 URL 추출


virustotal

바이러스토탈은 의심스런 파일과 URL을 분석하고 바이러스, 웜, 트로얀과 모든 종류의 악성 코드를 쉽고, 빠르게 탐지할 수 있는 편리한 무료 서비스입니다.

바이러스토탈 데이터셋을 통해 파일 검사 또는 검색을 사용해 보세요.

2. Virustotal을 이용해
추출한 URL 분석

elasticsearch

3. 분석 결과를
Elasticsearch에 저장&검색

SAMZO

[개발환경] 1. PyCharm

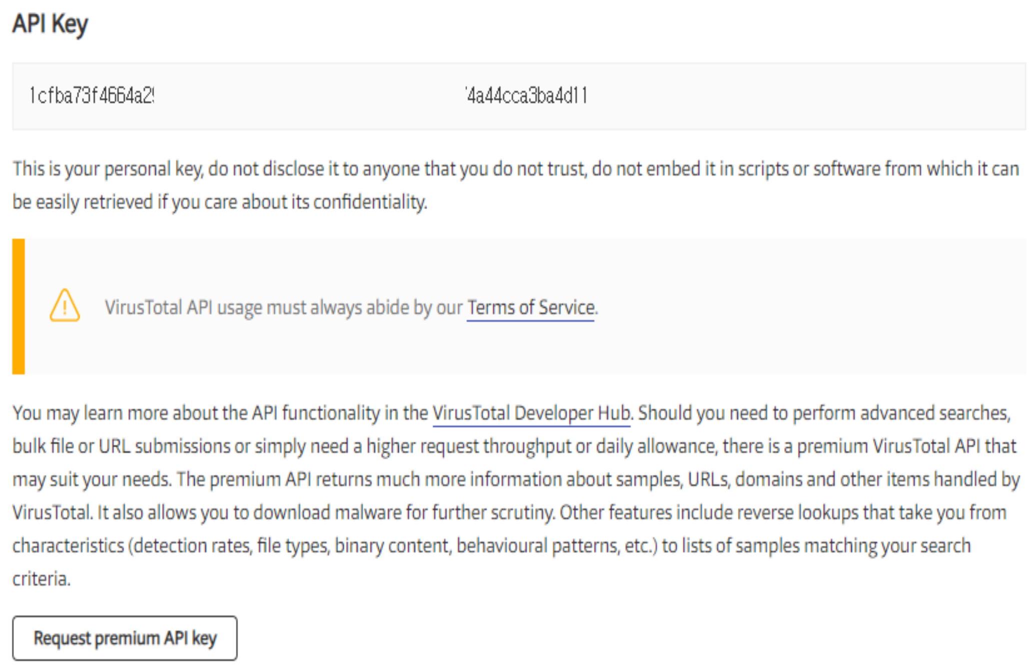
- 파이썬으로 개발할 때 사용하는 대표적인 도구
- <u>편리한 인터페이스</u>제공

→ 웹크롤링에 사용

SAMZO

[개발환경] 2. Virustotal

virustotal



API Key

1cfba73f4664a2   4a44cca3ba4d11

This is your personal key, do not disclose it to anyone that you do not trust, do not embed it in scripts or software from which it can be easily retrieved if you care about its confidentiality.

⚠ VirusTotal API usage must always abide by our Terms of Service.

You may learn more about the API functionality in the VirusTotal Developer Hub. Should you need to perform advanced searches, bulk file or URL submissions or simply need a higher request throughput or daily allowance, there is a premium VirusTotal API that may suit your needs. The premium API returns much more information about samples, URLs, domains and other items handled by VirusTotal. It also allows you to download malware for further scrutiny. Other features include reverse lookups that take you from characteristics (detection rates, file types, binary content, behavioural patterns, etc.) to lists of samples matching your search criteria.

- 의심스런 파일과 URL을 분석하고 바이러스, 웜, 트로얀과 모든 종류의 악성 코드를 쉽고, 빠르게 탐지할 수 있는 편리한 무료 서비스
- 여러 개의 백신 엔진으로 검사하여 그 결과를 투명하게 보여줌

→Virustotal API를 사용하여 크롤링한 URL 검사

[개발환경] 3. Elasticsearch

- 아파치 루신을 기반으로 만든 분산 검색엔진(오픈소스)
- 분산 시스템이기 때문에 검색 대상 용량이 증가했을 때 대처하기 쉬움
- 깃허브, 이베이, 가디언 같은 기업이 엘라스틱서치 기술로 내부 검색 기능을 구축함

- 데이터를 저장할 수 있어 NoSQL 저장소로도 활용할 수 있음
- 웹 서버나 데이터베이스 서버에 수정&삭제를 요청하고 결과를 받을 수 있는 응용 프로그래밍 인터페이스(API)를 지원함

→ Virustotal 분석 결과 저장&검색

[개발환경] 4. Kibana

- 엘라스틱 서치의 대시보드 어플리케이션
- Elasticsearch 데이터의 시각화와 Elastic Stack의 탐색을 지원

SAMZO

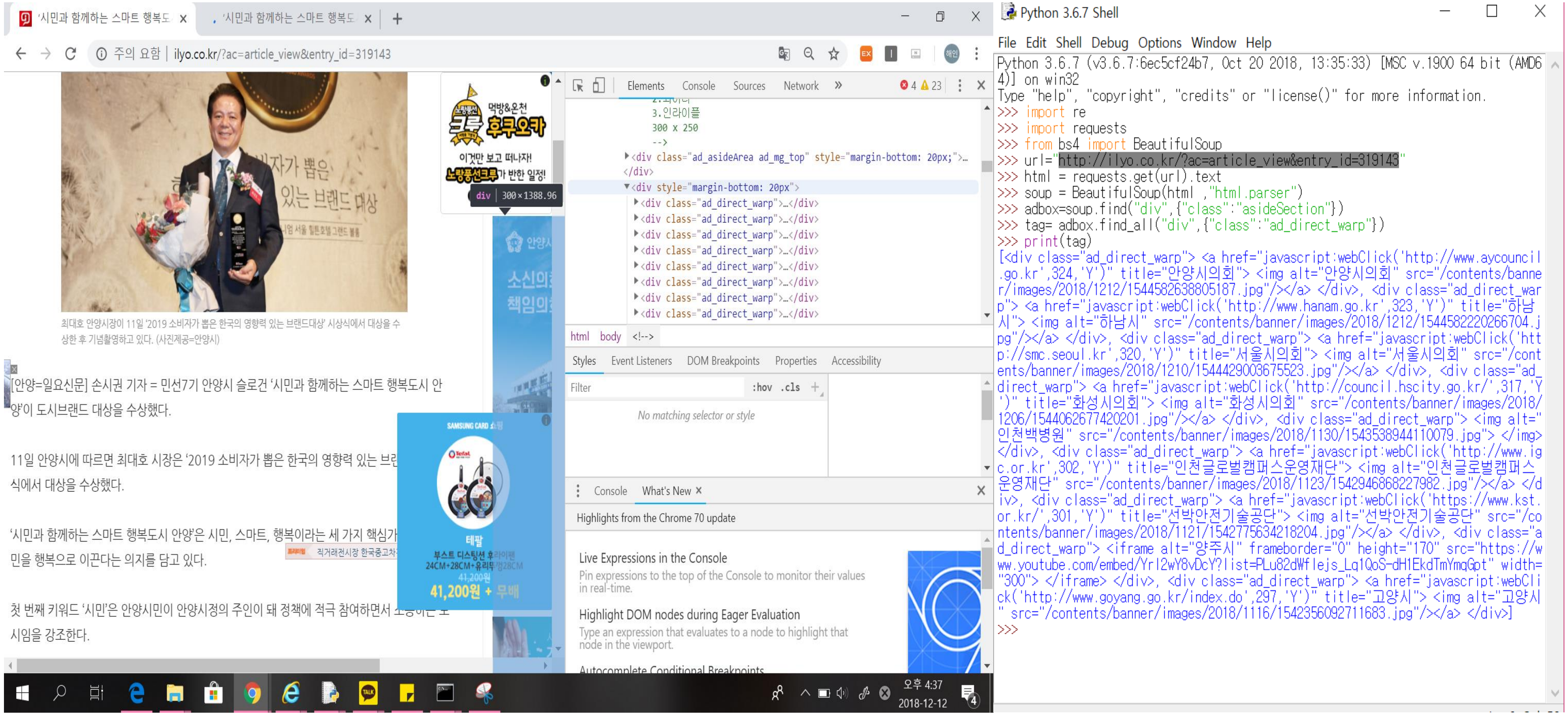
2. 구현

SAMZO

[웹크롤링]

SAMZO

[웹크롤링]



ilyo.co.kr/?ac=article_view&entry_id=319143

```
3.인라이플
300 x 250
-->
▶<div class="ad_asideArea ad_mg_top" style="margin-bottom: 20px;">…
</div>
▼<div style="margin-bottom: 20px">
  ▶<div class="ad_direct_warp">…</div>
  ▶<div class="ad_direct_warp">…</div>
  ▶<div class="ad_direct_warp">…</div>
  ▶<div class="ad_direct_warp">…</div>
  ▶<div class="ad_direct_warp">…</div>
  ▶<div class="ad_direct_warp">…</div>
  ▶<div class="ad_direct_warp">…</div>
  ▶<div class="ad_direct_warp">…</div>
```

최대호 안양시장이 11일 '2019 소비자가 뽑은 한국의 영향력 있는 브랜드대상' 시상식에서 대상을 수상한 후 기념촬영하고 있다. (사진제공=안양시)

[안양=일요신문] 손시권 기자 = 민선7기 안양시 슬로건 '시민과 함께하는 스마트 행복도시 안양'이 도시브랜드 대상을 수상했다.

11일 안양시에 따르면 최대호 시장은 '2019 소비자가 뽑은 한국의 영향력 있는 브랜
식에서 대상을 수상했다.

'시민과 함께하는 스마트 행복도시 안양'은 시민, 스마트, 행복이라는 세 가지 핵심가
민을 행복으로 이끈다는 의지를 담고 있다.

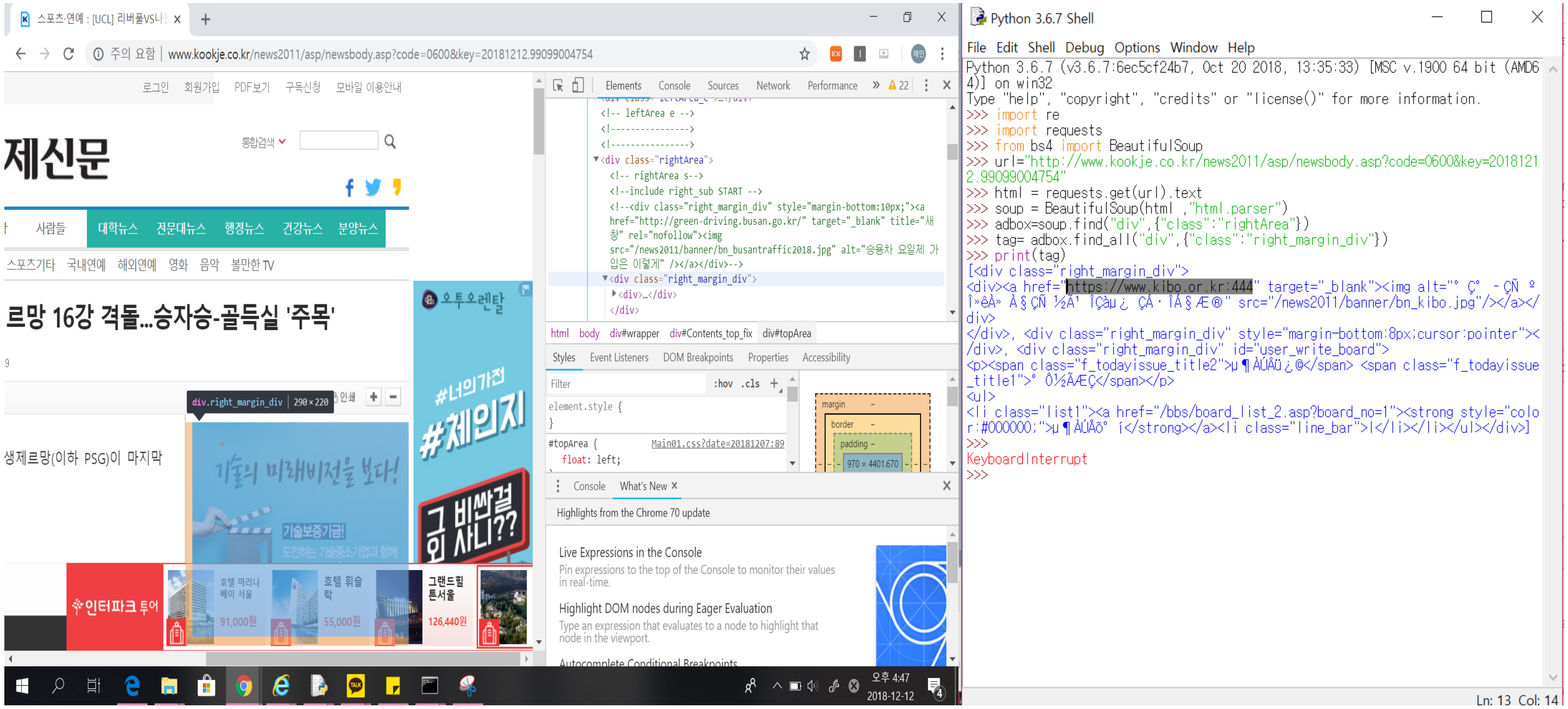
첫 번째 키워드 '시민'은 안양시민이 안양시정의 주인이 돼 정책에 적극 참여하면서 소통하는 도
시임을 강조한다.

Python 3.6.7 Shell

```
Python 3.6.7 (v3.6.7:6ec5cf24b7, Oct 20 2018, 13:35:33) [MSC v.1900 64 bit (AMD6
4)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> import re
>>> import requests
>>> from bs4 import BeautifulSoup
>>> url="http://ilyo.co.kr/?ac=article_view&entry_id=319143"
>>> html = requests.get(url).text
>>> soup = BeautifulSoup(html ,"html.parser")
>>> adbox=soup.find("div",{"class":"asideSection"})
>>> tag= adbox.find_all("div",{"class":"ad_direct_warp"})
>>> print(tag)
[<div class="ad_direct_warp"> <a href="javascript:webClick('http://www.aycouncil
.go.kr',324,'Y')" title="안양시의회"> <img alt="안양시의회" src="/contents/banne
r/images/2018/1212/1544582638805187.jpg"/></a> </div>, <div class="ad_direct_war
p"> <a href="javascript:webClick('http://www.hanam.go.kr',323,'Y')" title="하남
시"> <img alt="하남시" src="/contents/banner/images/2018/1212/154458220266704.j
pg"/></a> </div>, <div class="ad_direct_warp"> <a href="javascript:webClick('htt
p://smc.seoul.kr',320,'Y')" title="서울시의회"> <img alt="서울시의회" src="/cont
ents/banner/images/2018/1210/1544429003675523.jpg"/></a> </div>, <div class="ad_
direct_warp"> <a href="javascript:webClick('http://council.hscity.go.kr/',317,'Y
')" title="화성시의회"> <img alt="화성시의회" src="/contents/banner/images/2018/
1206/1544062677420201.jpg"/></a> </div>, <div class="ad_direct_warp"> <img alt="
인천백병원" src="/contents/banner/images/2018/1130/1543538944110079.jpg"> </img>
</div>, <div class="ad_direct_warp"> <a href="javascript:webClick('http://www.ig
c.or.kr',302,'Y')" title="인천글로벌캠퍼스운영재단"> <img alt="인천글로벌캠퍼스
운영재단" src="/contents/banner/images/2018/1123/1542946868227982.jpg"/></a> </d
iv>, <div class="ad_direct_warp"> <a href="javascript:webClick('https://www.kst.
or.kr/',301,'Y')" title="선박안전기술공단"> <img alt="선박안전기술공단" src="/co
ntents/banner/images/2018/1121/1542775634218204.jpg"/></a> </div>, <div class="a
d_direct_warp"> <iframe alt="양주시" frameborder="0" height="170" src="https://w
ww.youtube.com/embed/Yrl2wY8vDcY?list=PLu82dWflejs_Lq1QoS-dH1EkdTmYmqGpt" width=
"300"> </iframe> </div>, <div class="ad_direct_warp"> <a href="javascript:webCli
ck('http://www.goyang.go.kr/index.do',297,'Y')" title="고양시"> <img alt="고양시
" src="/contents/banner/images/2018/1116/1542356092711683.jpg"/></a> </div>]
>>>
```

SAMZO

[웹크롤링]


스포츠·연예 : [UCL] 리버풀VS나
www.kookje.co.kr/news2011/asp/newsbody.asp?code=0600&key=20181212.99099004754
로그인 회원가입 PDF보기 구독신청 모바일 이용안내
제신문
통합검색
사람들 대학뉴스 전문대뉴스 행정뉴스 건강뉴스 분양뉴스
스포츠기타 국내연예 해외연예 영화 음악 볼만한 TV
르망 16강 격돌…승자승-골득실 '주목'
div.right_margin_div | 290 × 220
생제르망(이하 PSG)이 마지막
Elements Console Sources Network Performance
html body div#wrapper div#Contents_top_fix div#topArea
Styles Event Listeners DOM Breakpoints Properties Accessibility
margin border padding 970 × 4401.670
Console What's New
Highlights from the Chrome 70 update
Live Expressions in the Console
Pin expressions to the top of the Console to monitor their values in real-time.
Highlight DOM nodes during Eager Evaluation
Type an expression that evaluates to a node to highlight that node in the viewport.
Autocomplete Conditional Breakpoints
오후 4:47
2018-12-12
Python 3.6.7 Shell
File Edit Shell Debug Options Window Help
Ln: 13 Col: 14

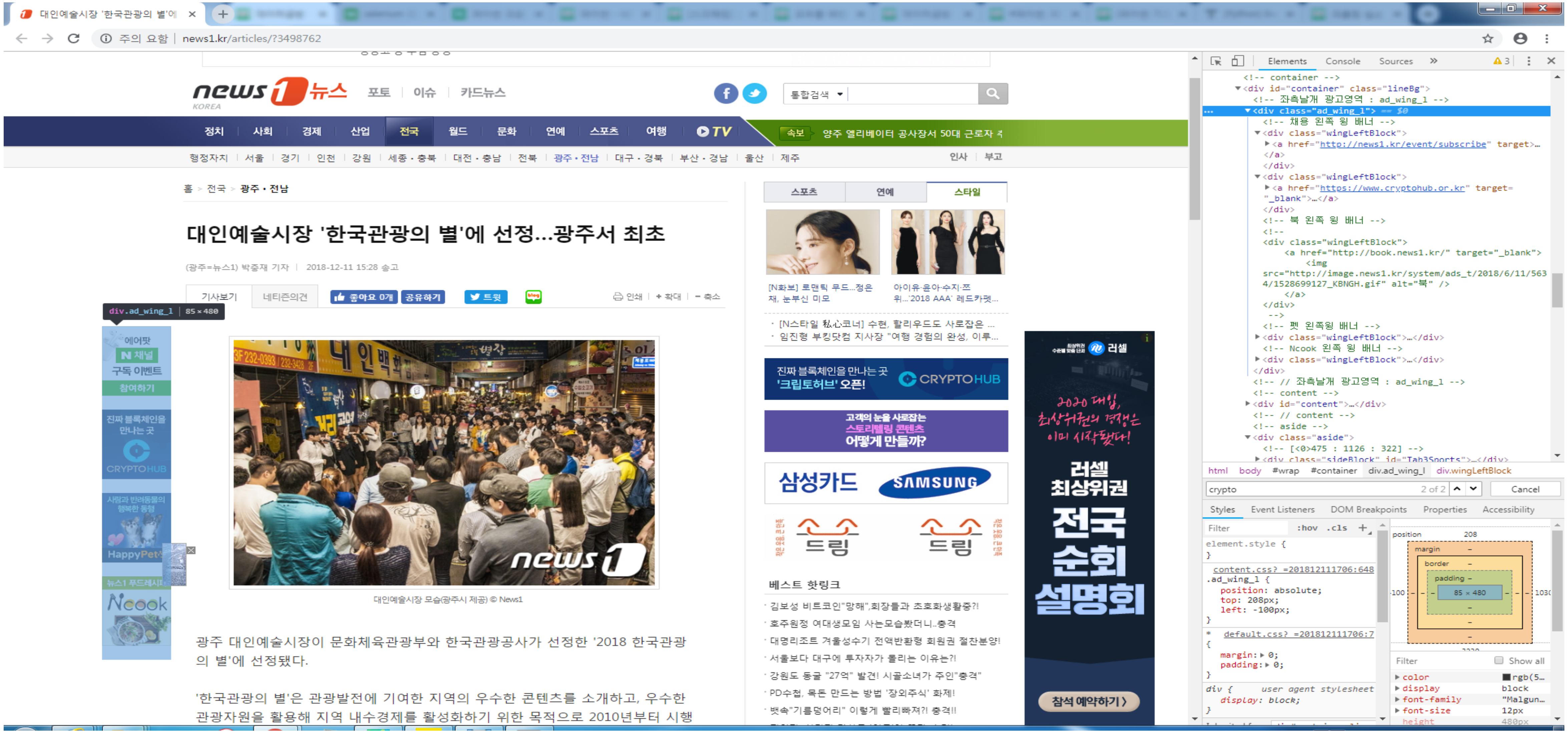
```
<!-- leftArea e -->
<!--------------->
<!--------------->
▼<div class="rightArea">
  <!-- rightArea s-->
  <!--include right_sub START -->
  <!--<div class="right_margin_div" style="margin-bottom:10px;"><a
  href="http://green-driving.busan.go.kr/" target="_blank" title="새
  창" rel="nofollow"><img
  src="/news2011/banner/bn_busantraffic2018.jpg" alt="승용차 요일제 가
  입은 이렇게" /></a></div>-->
 ▼<div class="right_margin_div">
   ▶<div>…</div>
  </div>
```

```
element.style {
}
#topArea {                    Main01.css?date=20181207:89
  float: left;
```

```
Python 3.6.7 (v3.6.7:6ec5cf24b7, Oct 20 2018, 13:35:33) [MSC v.1900 64 bit (AMD6
4)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>> import re
>>> import requests
>>> from bs4 import BeautifulSoup
>>> url="http://www.kookje.co.kr/news2011/asp/newsbody.asp?code=0600&key=2018121
2.99099004754"
>>> html = requests.get(url).text
>>> soup = BeautifulSoup(html ,"html.parser")
>>> adbox=soup.find("div",{"class":"rightArea"})
>>> tag= adbox.find_all("div",{"class":"right_margin_div"})
>>> print(tag)
[<div class="right_margin_div">
<div><a href="https://www.kibo.or.kr:444" target="_blank"><img alt="° Ç°  - ÇÑ º
Î»êÀ» À§ÇÑ ½Å¹  ÎÇàµ¿  ÇÀ· ÎÀ§Æ®" src="/news2011/banner/bn_kibo.jpg"/></a></
div>
</div>, <div class="right_margin_div" style="margin-bottom:8px;cursor:pointer"><
/div>, <div class="right_margin_div" id="user_write_board">
<p><span class="f_todayissue_title2">µ¶ÀÚÂü¿©</span> <span class="f_todayissue
_title1">° Ô½ÃÆÇ</span></p>
<ul>
<li class="list1"><a href="/bbs/board_list_2.asp?board_no=1"><strong style="colo
r:#000000;">µ¶ÀÚÅõ° í</strong></a><li class="line_bar">|</li></li></ul></div>]
>>>
KeyboardInterrupt
>>>
```

SAMZO

[웹크롤링]



news1.kr/articles/?3498762

홈 > 전국 > 광주 · 전남

대인예술시장 '한국관광의 별'에 선정…광주서 최초

(광주=뉴스1) 박중재 기자 | 2018-12-11 15:28 송고

대인예술시장 모습(광주시 제공) © News1

광주 대인예술시장이 문화체육관광부와 한국관광공사가 선정한 '2018 한국관광의 별'에 선정됐다.

'한국관광의 별'은 관광발전에 기여한 지역의 우수한 콘텐츠를 소개하고, 우수한 관광자원을 활용해 지역 내수경제를 활성화하기 위한 목적으로 2010년부터 시행

div.ad_wing_l | 85 × 480

```
<!-- container -->
▼<div id="container" class="lineBg">
   <!-- 좌측날개 광고영역 : ad_wing_l -->
 ▼<div class="ad_wing_l"> == $0
     <!-- 채용 왼쪽 윙 배너 -->
   ▼<div class="wingLeftBlock">
     ▶<a href="http://news1.kr/event/subscribe" target>…
     </a>
     </div>
   ▼<div class="wingLeftBlock">
     ▶<a href="https://www.cryptohub.or.kr" target=
     "_blank">…</a>
     </div>
     <!-- 북 왼쪽 윙 배너 -->
     <!--
     <div class="wingLeftBlock">
         <a href="http://book.news1.kr/" target="_blank">
             <img
     src="http://image.news1.kr/system/ads_t/2018/6/11/563
     4/1528699127_KBNGH.gif" alt="북" />
         </a>
     </div>
      -->
     <!-- 펫 왼쪽윙 배너 -->
   ▶<div class="wingLeftBlock">…</div>
     <!-- Ncook 왼쪽 윙 배너 -->
   ▶<div class="wingLeftBlock">…</div>
   </div>
   <!-- // 좌측날개 광고영역 : ad_wing_l -->
   <!-- content -->
 ▶<div id="content">…</div>
   <!-- // content -->
   <!-- aside -->
 ▼<div class="aside">
     <!-- [<0>475 : 1126 : 322] -->
   ▶<div class="sideBlock" id="Tab3Sports">…</div>
```

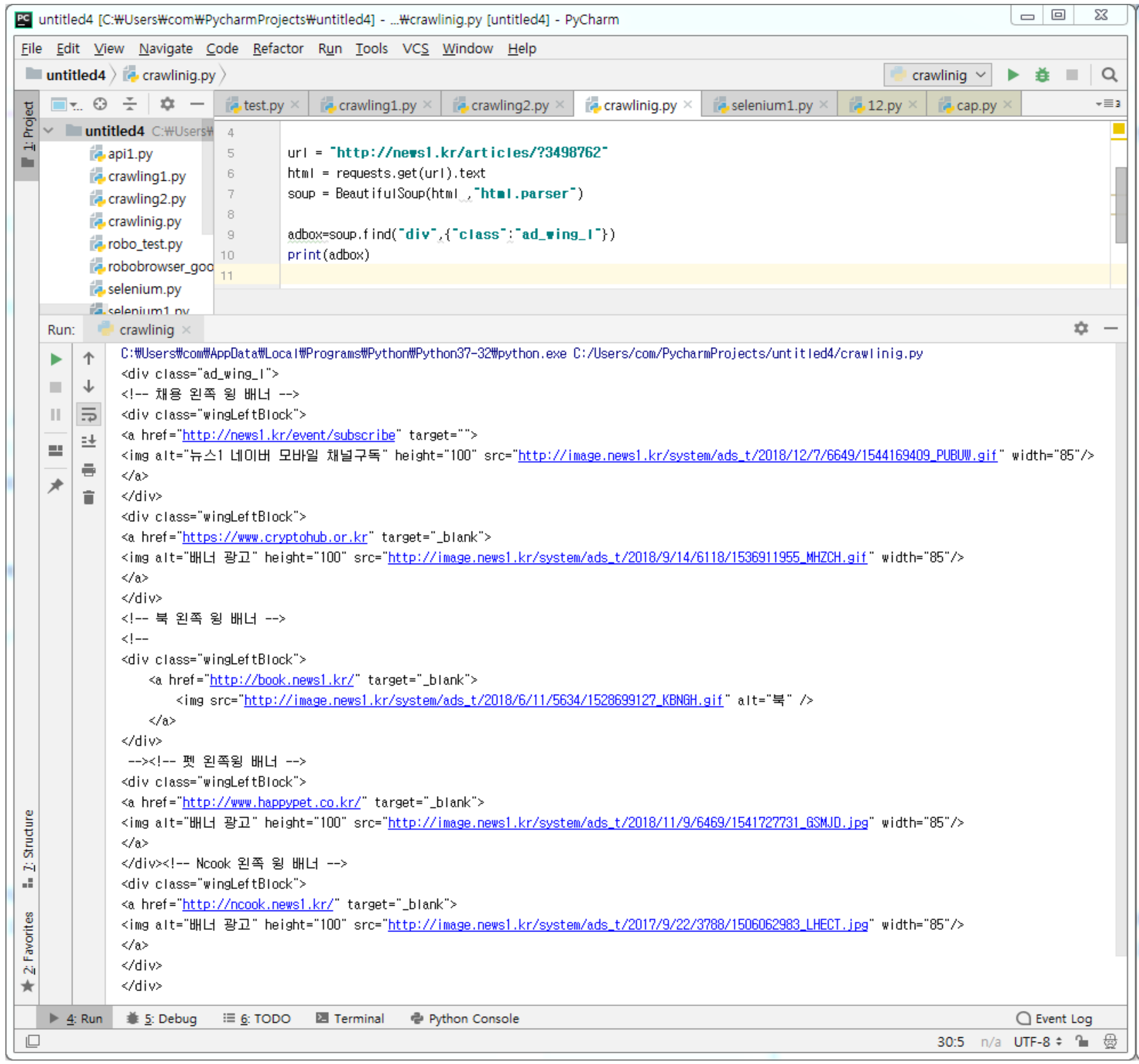
html body #wrap #container div.ad_wing_l div.wingLeftBlock

crypto 2 of 2

```
element.style {
}
content.css?_=201812111706:648
.ad_wing_l {
  position: absolute;
  top: 208px;
  left: -100px;
}
* default.css?_=201812111706:7
{
  margin:▶ 0;
  padding:▶ 0;
}
div {        user agent stylesheet
  display: block;
}
```

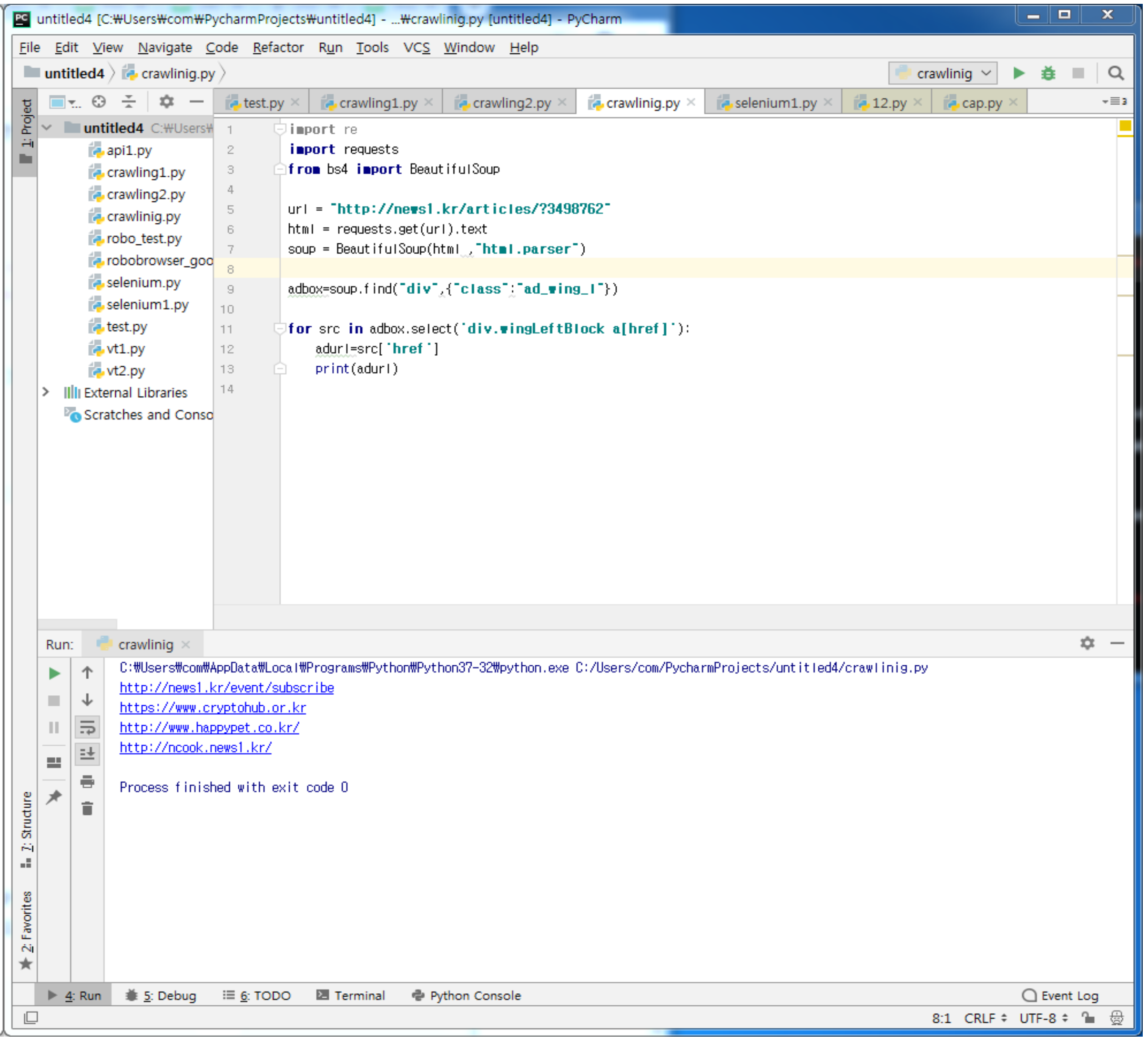
```
position 208
margin –
border –
padding –
-100  85 × 480  1030

color        rgb(5…
display      block
font-family  "Malgun…
font-size    12px
height       480px
```

[웹크롤링]



```python
url = "http://news1.kr/articles/?3498762"
html = requests.get(url).text
soup = BeautifulSoup(html ,"html.parser")

adbox=soup.find("div",{"class":"ad_wing_l"})
print(adbox)
```

```
C:₩Users₩com₩AppData₩Local₩Programs₩Python₩Python37-32₩python.exe C:/Users/com/PycharmProjects/untitled4/crawlinig.py
<div class="ad_wing_l">
<!-- 채용 왼쪽 윙 배너 -->
<div class="wingLeftBlock">
<a href="http://news1.kr/event/subscribe" target="">
<img alt="뉴스1 네이버 모바일 채널구독" height="100" src="http://image.news1.kr/system/ads_t/2018/12/7/6649/1544169409_PUBUW.gif" width="85"/>
</a>
</div>
<div class="wingLeftBlock">
<a href="https://www.cryptohub.or.kr" target="_blank">
<img alt="배너 광고" height="100" src="http://image.news1.kr/system/ads_t/2018/9/14/6118/1536911955_MHZCH.gif" width="85"/>
</a>
</div>
<!-- 북 왼쪽 윙 배너 -->
<!--
<div class="wingLeftBlock">
    <a href="http://book.news1.kr/" target="_blank">
        <img src="http://image.news1.kr/system/ads_t/2018/6/11/5634/1528699127_KBNGH.gif" alt="북" />
    </a>
</div>
 --><!-- 펫 왼쪽윙 배너 -->
<div class="wingLeftBlock">
<a href="http://www.happypet.co.kr/" target="_blank">
<img alt="배너 광고" height="100" src="http://image.news1.kr/system/ads_t/2018/11/9/6469/1541727731_GSMJD.jpg" width="85"/>
</a>
</div><!-- Ncook 왼쪽 윙 배너 -->
<div class="wingLeftBlock">
<a href="http://ncook.news1.kr/" target="_blank">
<img alt="배너 광고" height="100" src="http://image.news1.kr/system/ads_t/2017/9/22/3788/1506062983_LHECT.jpg" width="85"/>
</a>
</div>
</div>
```



```python
import re
import requests
from bs4 import BeautifulSoup

url = "http://news1.kr/articles/?3498762"
html = requests.get(url).text
soup = BeautifulSoup(html ,"html.parser")

adbox=soup.find("div",{"class":"ad_wing_l"})

for src in adbox.select('div.wingLeftBlock a[href]'):
    adurl=src['href']
    print(adurl)
```

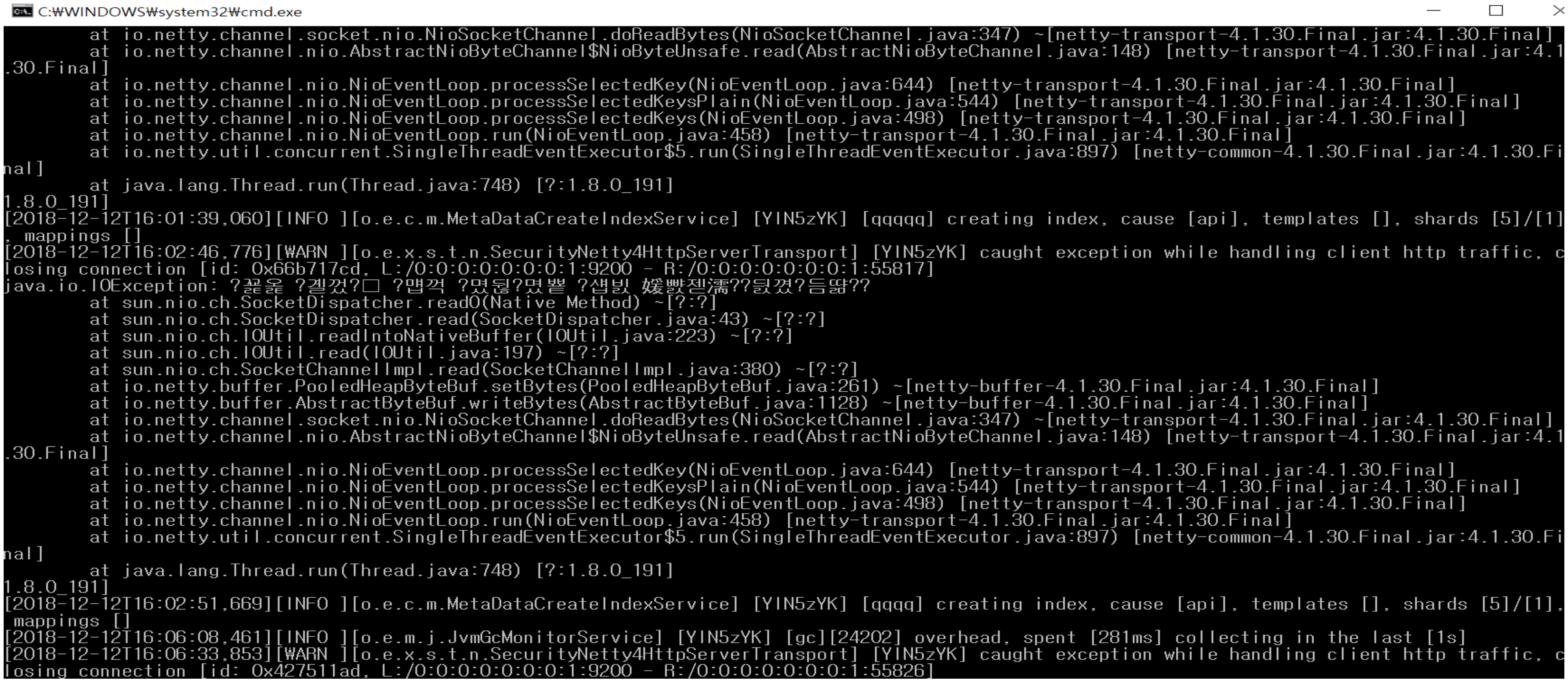
```
C:₩Users₩com₩AppData₩Local₩Programs₩Python₩Python37-32₩python.exe C:/Users/com/PycharmProjects/untitled4/crawlinig.py
http://news1.kr/event/subscribe
https://www.cryptohub.or.kr
http://www.happypet.co.kr/
http://ncook.news1.kr/

Process finished with exit code 0
```

SAMZO

[윈도우 환경에서 엘라스틱서치 DB 사용]

[윈도우 환경에서 엘라스틱서치 DB 사용]

Elasticsearch를 다운 받은 후, elasticsearch.bat을 실행시킨다.



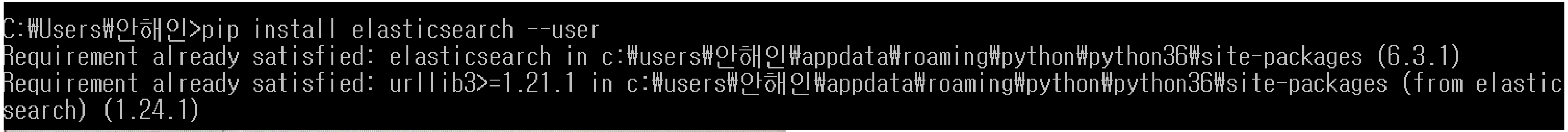
C:￦WINDOWS￦system32￦cmd.exe

```
        at io.netty.channel.socket.nio.NioSocketChannel.doReadBytes(NioSocketChannel.java:347) ~[netty-transport-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.channel.nio.AbstractNioByteChannel$NioByteUnsafe.read(AbstractNioByteChannel.java:148) [netty-transport-4.1.30.Final.jar:4.1
.30.Final]
        at io.netty.channel.nio.NioEventLoop.processSelectedKey(NioEventLoop.java:644) [netty-transport-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.channel.nio.NioEventLoop.processSelectedKeysPlain(NioEventLoop.java:544) [netty-transport-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.channel.nio.NioEventLoop.processSelectedKeys(NioEventLoop.java:498) [netty-transport-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:458) [netty-transport-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.util.concurrent.SingleThreadEventExecutor$5.run(SingleThreadEventExecutor.java:897) [netty-common-4.1.30.Final.jar:4.1.30.Fi
nal]
        at java.lang.Thread.run(Thread.java:748) [?:1.8.0_191]
1.8.0_191]
[2018-12-12T16:01:39,060][INFO ][o.e.c.m.MetaDataCreateIndexService] [YIN5zYK] [qqqqq] creating index, cause [api], templates [], shards [5]/[1]
, mappings []
[2018-12-12T16:02:46,776][WARN ][o.e.x.s.t.n.SecurityNetty4HttpServerTransport] [YIN5zYK] caught exception while handling client http traffic, c
losing connection [id: 0x66b717cd, L:/0:0:0:0:0:0:0:1:9200 - R:/0:0:0:0:0:0:0:1:55817]
java.io.IOException: ?꿡욌 ?겢겠?□ ?맵걱 ?몄뒿?몃뿉 ?샙빐 媛뺤젣濡??뒿겼?듬땅??
        at sun.nio.ch.SocketDispatcher.read0(Native Method) ~[?:?]
        at sun.nio.ch.SocketDispatcher.read(SocketDispatcher.java:43) ~[?:?]
        at sun.nio.ch.IOUtil.readIntoNativeBuffer(IOUtil.java:223) ~[?:?]
        at sun.nio.ch.IOUtil.read(IOUtil.java:197) ~[?:?]
        at sun.nio.ch.SocketChannelImpl.read(SocketChannelImpl.java:380) ~[?:?]
        at io.netty.buffer.PooledHeapByteBuf.setBytes(PooledHeapByteBuf.java:261) ~[netty-buffer-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.buffer.AbstractByteBuf.writeBytes(AbstractByteBuf.java:1128) ~[netty-buffer-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.channel.socket.nio.NioSocketChannel.doReadBytes(NioSocketChannel.java:347) ~[netty-transport-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.channel.nio.AbstractNioByteChannel$NioByteUnsafe.read(AbstractNioByteChannel.java:148) [netty-transport-4.1.30.Final.jar:4.1
.30.Final]
        at io.netty.channel.nio.NioEventLoop.processSelectedKey(NioEventLoop.java:644) [netty-transport-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.channel.nio.NioEventLoop.processSelectedKeysPlain(NioEventLoop.java:544) [netty-transport-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.channel.nio.NioEventLoop.processSelectedKeys(NioEventLoop.java:498) [netty-transport-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:458) [netty-transport-4.1.30.Final.jar:4.1.30.Final]
        at io.netty.util.concurrent.SingleThreadEventExecutor$5.run(SingleThreadEventExecutor.java:897) [netty-common-4.1.30.Final.jar:4.1.30.Fi
nal]
        at java.lang.Thread.run(Thread.java:748) [?:1.8.0_191]
1.8.0_191]
[2018-12-12T16:02:51,669][INFO ][o.e.c.m.MetaDataCreateIndexService] [YIN5zYK] [qqqq] creating index, cause [api], templates [], shards [5]/[1],
 mappings []
[2018-12-12T16:06:08,461][INFO ][o.e.m.j.JvmGcMonitorService] [YIN5zYK] [gc][24202] overhead, spent [281ms] collecting in the last [1s]
[2018-12-12T16:06:33,853][WARN ][o.e.x.s.t.n.SecurityNetty4HttpServerTransport] [YIN5zYK] caught exception while handling client http traffic, c
losing connection [id: 0x427511ad, L:/0:0:0:0:0:0:0:1:9200 - R:/0:0:0:0:0:0:0:1:55826]
```

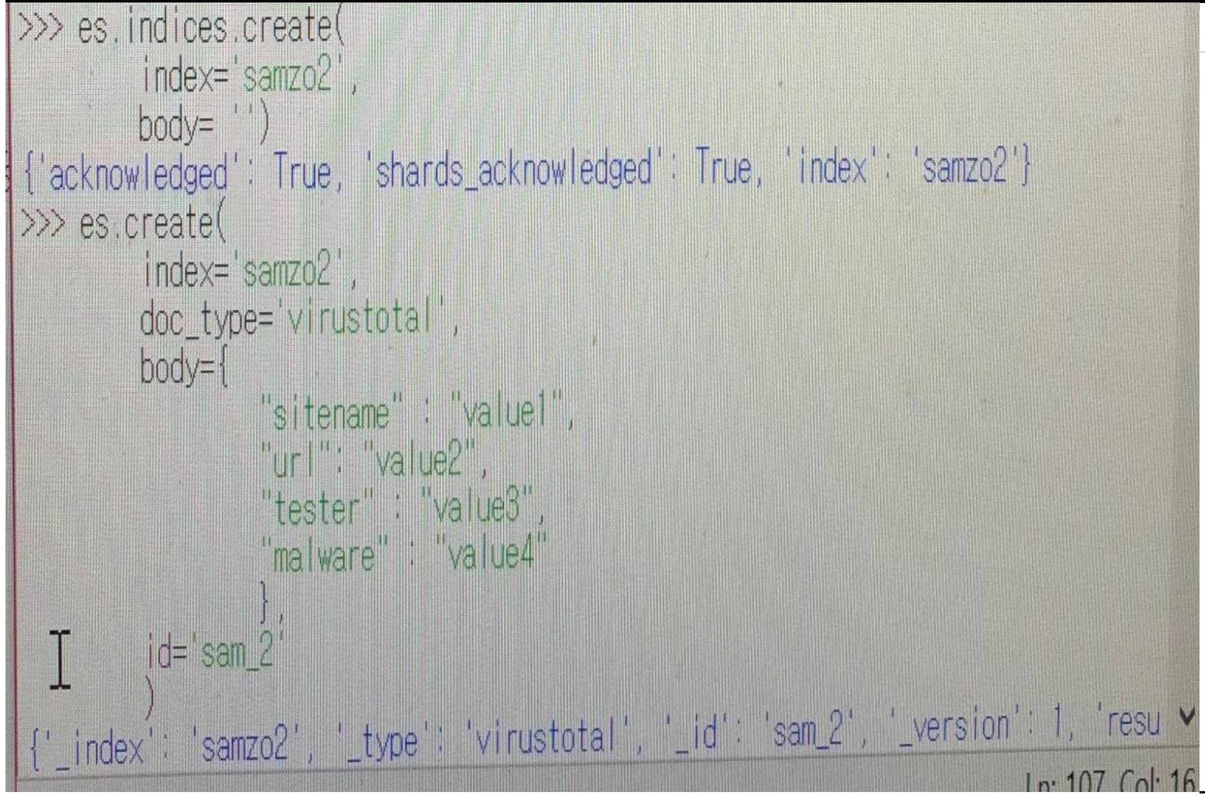
SAMZO

[윈도우 환경에서 엘라스틱서치 DB 사용]

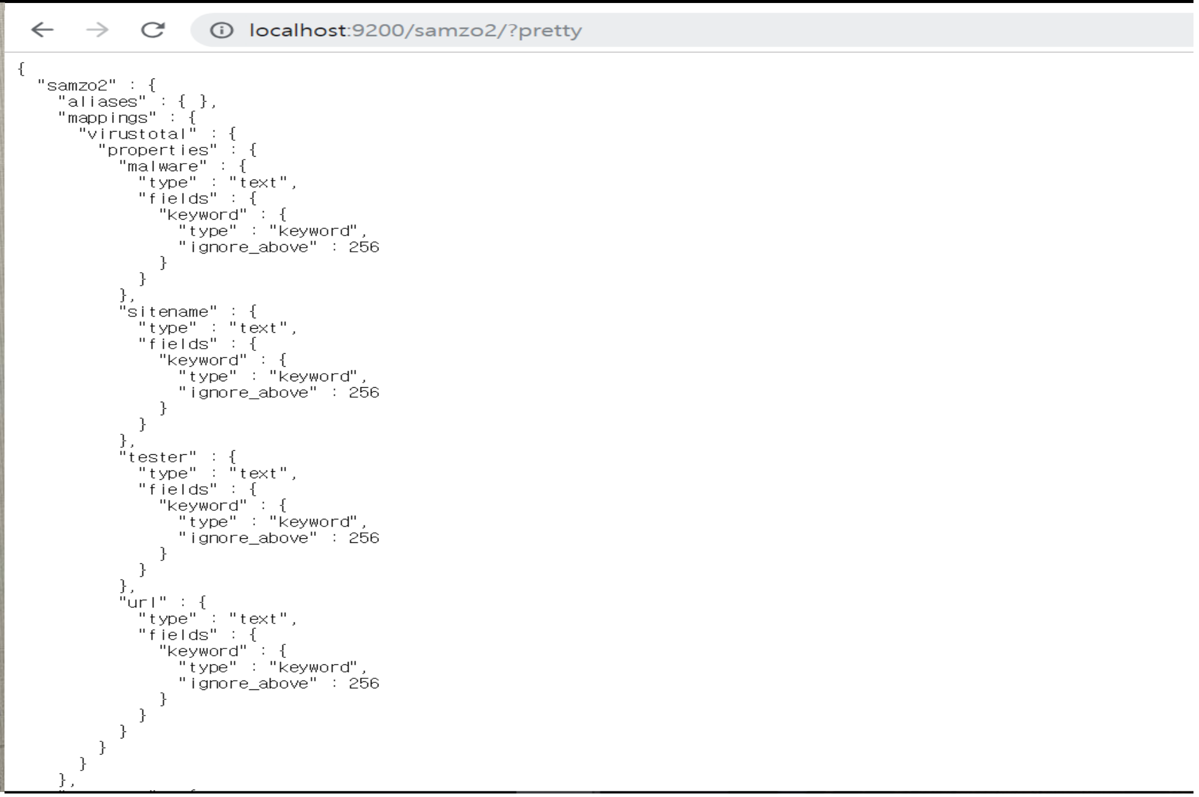
Python에서 elasticsearch를 사용하기 위해 pip install을 한 후, 인덱스 생성



```
C:₩Users₩안해인>pip install elasticsearch --user
Requirement already satisfied: elasticsearch in c:₩users₩안해인₩appdata₩roaming₩python₩python36₩site-packages (6.3.1)
Requirement already satisfied: urllib3>=1.21.1 in c:₩users₩안해인₩appdata₩roaming₩python₩python36₩site-packages (from elastic
search) (1.24.1)
```



```
>>> es.indices.create(
        index='samzo2',
        body='')
{'acknowledged': True, 'shards_acknowledged': True, 'index': 'samzo2'}
>>> es.create(
        index='samzo2',
        doc_type='virustotal',
        body={
                "sitename" : "value1",
                "url": "value2",
                "tester" : "value3",
                "malware" : "value4"
                },
        id='sam_2'
        )
{'_index': 'samzo2', '_type': 'virustotal', '_id': 'sam_2', '_version': 1, 'resu
                                                                    Ln: 107 Col: 16
```



localhost:9200/samzo2/?pretty

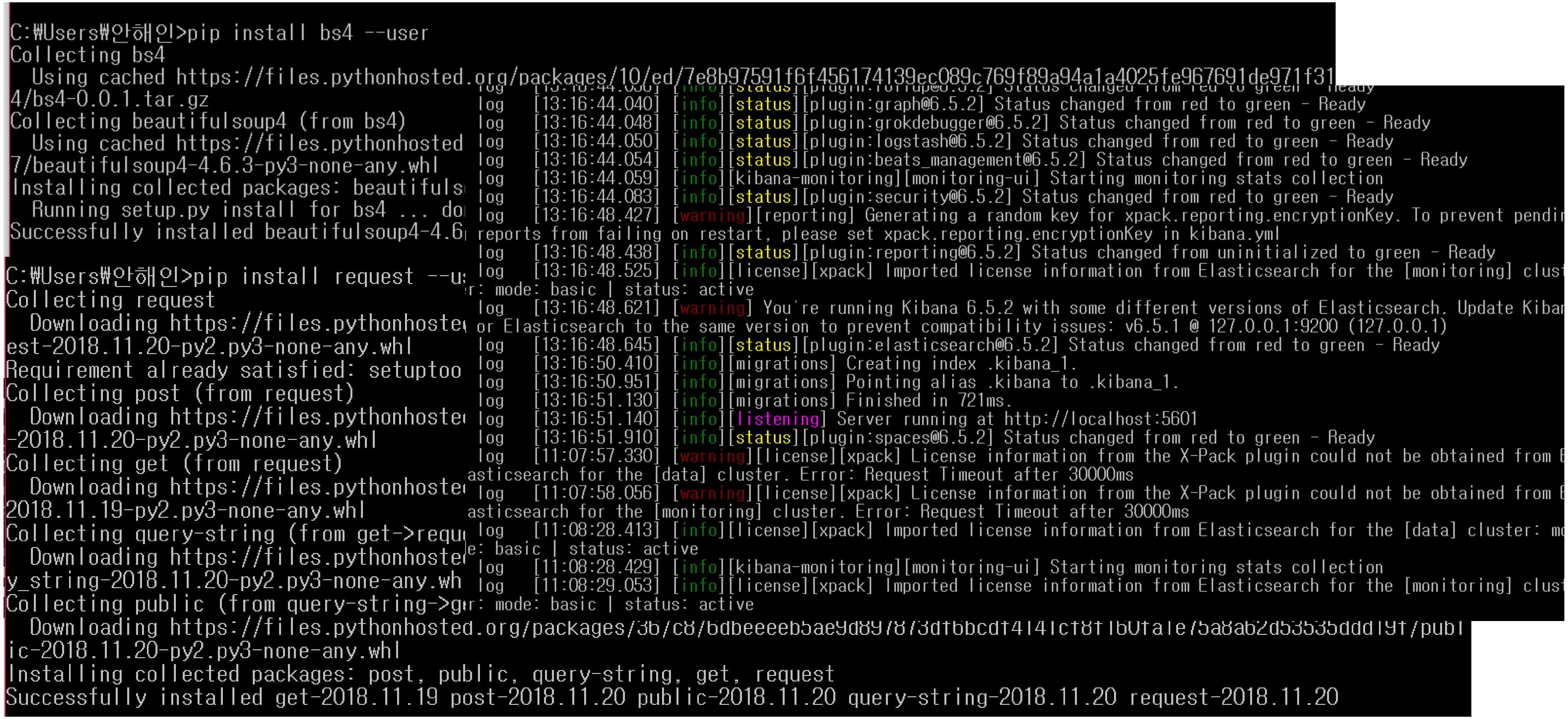
```
{
  "samzo2" : {
    "aliases" : { },
    "mappings" : {
      "virustotal" : {
        "properties" : {
          "malware" : {
            "type" : "text",
            "fields" : {
              "keyword" : {
                "type" : "keyword",
                "ignore_above" : 256
              }
            }
          },
          "sitename" : {
            "type" : "text",
            "fields" : {
              "keyword" : {
                "type" : "keyword",
                "ignore_above" : 256
              }
            }
          },
          "tester" : {
            "type" : "text",
            "fields" : {
              "keyword" : {
                "type" : "keyword",
                "ignore_above" : 256
              }
            }
          },
          "url" : {
            "type" : "text",
            "fields" : {
              "keyword" : {
                "type" : "keyword",
                "ignore_above" : 256
              }
            }
          }
        }
      }
    },
```

SAMZO

[바이러스토탈]

SAMZO

[바이러스토탈]



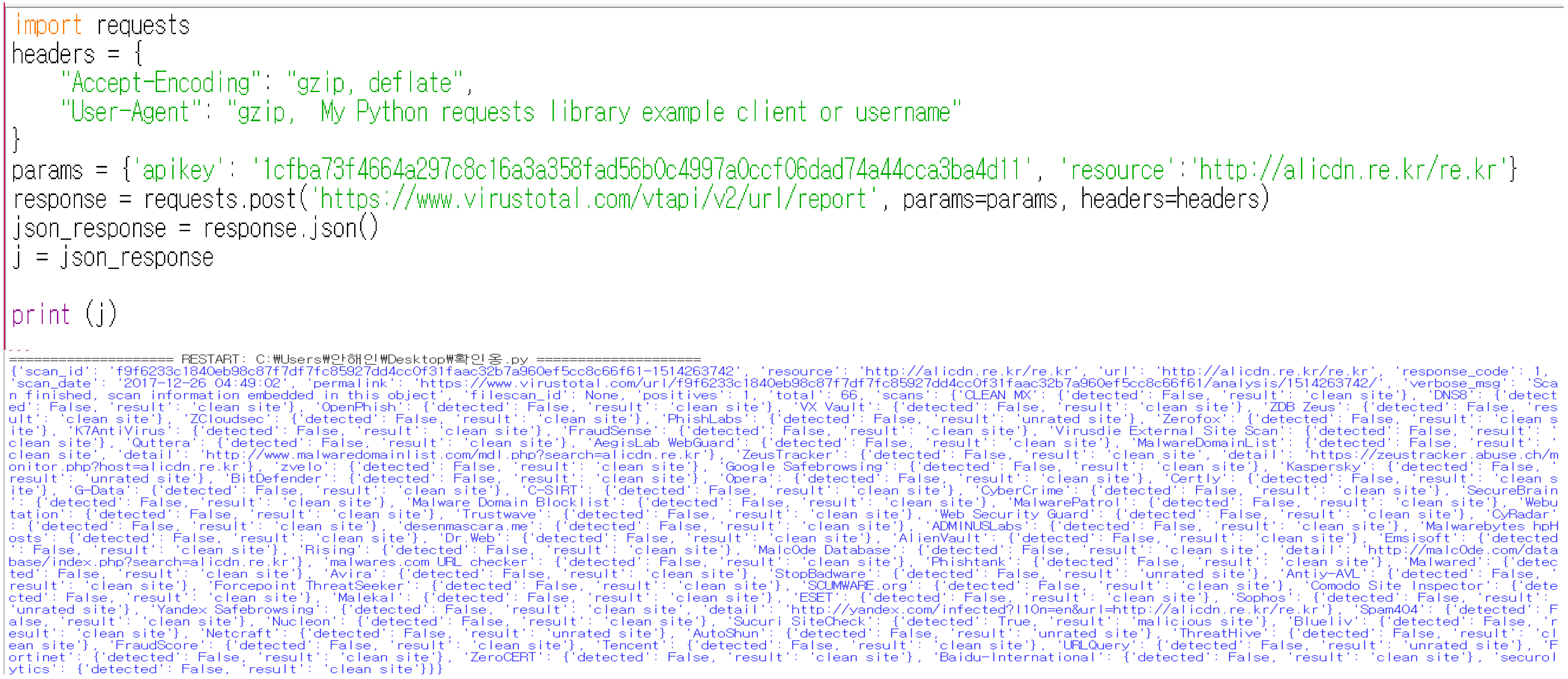
```
C:₩Users₩안해인>pip install bs4 --user
Collecting bs4
  Using cached https://files.pythonhosted.org/packages/10/ed/7e8b97591f6f456174139ec089c769f89a94a1a4025fe967691de971f31
4/bs4-0.0.1.tar.gz
Collecting beautifulsoup4 (from bs4)
  Using cached https://files.pythonhosted
7/beautifulsoup4-4.6.3-py3-none-any.whl
Installing collected packages: beautifuls
  Running setup.py install for bs4 ... do
Successfully installed beautifulsoup4-4.6

C:₩Users₩안해인>pip install request --u
Collecting request
  Downloading https://files.pythonhoste
est-2018.11.20-py2.py3-none-any.whl
Requirement already satisfied: setuptoo
Collecting post (from request)
  Downloading https://files.pythonhoste
-2018.11.20-py2.py3-none-any.whl
Collecting get (from request)
  Downloading https://files.pythonhoste
2018.11.19-py2.py3-none-any.whl
Collecting query-string (from get->requ
  Downloading https://files.pythonhoste
y_string-2018.11.20-py2.py3-none-any.wh
Collecting public (from query-string->g
  Downloading https://files.pythonhosted.org/packages/3b/c8/bdbeeeeb5ae9d89787 3dfbbcdf4141cf8f160fa1e75a8ab2d53535ddd19f/publ
ic-2018.11.20-py2.py3-none-any.whl
Installing collected packages: post, public, query-string, get, request
Successfully installed get-2018.11.19 post-2018.11.20 public-2018.11.20 query-string-2018.11.20 request-2018.11.20
```

```
  log   [13:16:44.040] [info][status][plugin:graph@6.5.2] Status changed from red to green - Ready
  log   [13:16:44.048] [info][status][plugin:grokdebugger@6.5.2] Status changed from red to green - Ready
  log   [13:16:44.050] [info][status][plugin:logstash@6.5.2] Status changed from red to green - Ready
  log   [13:16:44.054] [info][status][plugin:beats_management@6.5.2] Status changed from red to green - Ready
  log   [13:16:44.059] [info][kibana-monitoring][monitoring-ui] Starting monitoring stats collection
  log   [13:16:44.083] [info][status][plugin:security@6.5.2] Status changed from red to green - Ready
  log   [13:16:48.427] [warning][reporting] Generating a random key for xpack.reporting.encryptionKey. To prevent pendi
reports from failing on restart, please set xpack.reporting.encryptionKey in kibana.yml
  log   [13:16:48.438] [info][status][plugin:reporting@6.5.2] Status changed from uninitialized to green - Ready
  log   [13:16:48.525] [info][license][xpack] Imported license information from Elasticsearch for the [monitoring] clus
r: mode: basic | status: active
  log   [13:16:48.621] [warning] You're running Kibana 6.5.2 with some different versions of Elasticsearch. Update Kiba
or Elasticsearch to the same version to prevent compatibility issues: v6.5.1 @ 127.0.0.1:9200 (127.0.0.1)
  log   [13:16:48.645] [info][status][plugin:elasticsearch@6.5.2] Status changed from red to green - Ready
  log   [13:16:50.410] [info][migrations] Creating index .kibana_1.
  log   [13:16:50.951] [info][migrations] Pointing alias .kibana to .kibana_1.
  log   [13:16:51.130] [info][migrations] Finished in 721ms.
  log   [13:16:51.140] [info][listening] Server running at http://localhost:5601
  log   [13:16:51.910] [info][status][plugin:spaces@6.5.2] Status changed from red to green - Ready
  log   [11:07:57.330] [warning][license][xpack] License information from the X-Pack plugin could not be obtained from E
asticsearch for the [data] cluster. Error: Request Timeout after 30000ms
  log   [11:07:58.056] [warning][license][xpack] License information from the X-Pack plugin could not be obtained from E
asticsearch for the [monitoring] cluster. Error: Request Timeout after 30000ms
  log   [11:08:28.413] [info][license][xpack] Imported license information from Elasticsearch for the [data] cluster: m
e: basic | status: active
  log   [11:08:28.429] [info][kibana-monitoring][monitoring-ui] Starting monitoring stats collection
  log   [11:08:29.053] [info][license][xpack] Imported license information from Elasticsearch for the [monitoring] clus
r: mode: basic | status: active
```

SAMZO

[바이러스토탈]

해당 URL을 바이러스토탈로 분석 후 나온 결과값을 출력



```python
import requests
headers = {
    "Accept-Encoding": "gzip, deflate",
    "User-Agent": "gzip,  My Python requests library example client or username"
}
params = {'apikey': '1cfba73f4664a297c8c16a3a358fad56b0c4997a0ccf06dad74a44cca3ba4d11', 'resource':'http://alicdn.re.kr/re.kr'}
response = requests.post('https://www.virustotal.com/vtapi/v2/url/report', params=params, headers=headers)
json_response = response.json()
j = json_response

print (j)
```

```
==================== RESTART: C:₩Users₩안해인₩Desktop₩확인용.py ====================
{'scan_id': 'f9f6233c1840eb98c87f7df7fc85927dd4cc0f31faac32b7a960ef5cc8c66f61-1514263742', 'resource': 'http://alicdn.re.kr/re.kr', 'url': 'http://alicdn.re.kr/re.kr', 'response_code': 1,
'scan_date': '2017-12-26 04:49:02', 'permalink': 'https://www.virustotal.com/url/f9f6233c1840eb98c87f7df7fc85927dd4cc0f31faac32b7a960ef5cc8c66f61/analysis/1514263742/', 'verbose_msg': 'Sca
n finished, scan information embedded in this object', 'filescan_id': None, 'positives': 1, 'total': 66, 'scans': {'CLEAN MX': {'detected': False, 'result': 'clean site'}, 'DNS8': {'detect
ed': False, 'result': 'clean site'}, 'OpenPhish': {'detected': False, 'result': 'clean site'}, 'VX Vault': {'detected': False, 'result': 'clean site'}, 'ZDB Zeus': {'detected': False, 'res
ult': 'clean site'}, 'ZCloudsec': {'detected': False, 'result': 'clean site'}, 'PhishLabs': {'detected': False, 'result': 'unrated site'}, 'Zerofox': {'detected': False, 'result': 'clean s
ite'}, 'K7AntiVirus': {'detected': False, 'result': 'clean site'}, 'FraudSense': {'detected': False, 'result': 'clean site'}, 'Virusdie External Site Scan': {'detected': False, 'result': '
clean site'}, 'Quttera': {'detected': False, 'result': 'clean site'}, 'AegisLab WebGuard': {'detected': False, 'result': 'clean site'}, 'MalwareDomainList': {'detected': False, 'result': '
clean site', 'detail': 'http://www.malwaredomainlist.com/mdl.php?search=alicdn.re.kr'}, 'ZeusTracker': {'detected': False, 'result': 'clean site', 'detail': 'https://zeustracker.abuse.ch/m
onitor.php?host=alicdn.re.kr'}, 'zvelo': {'detected': False, 'result': 'clean site'}, 'Google Safebrowsing': {'detected': False, 'result': 'clean site'}, 'Kaspersky': {'detected': False, '
result': 'unrated site'}, 'BitDefender': {'detected': False, 'result': 'clean site'}, 'Opera': {'detected': False, 'result': 'clean site'}, 'Certly': {'detected': False, 'result': 'clean s
ite'}, 'G-Data': {'detected': False, 'result': 'clean site'}, 'C-SIRT': {'detected': False, 'result': 'clean site'}, 'CyberCrime': {'detected': False, 'result': 'clean site'}, 'SecureBrain
': {'detected': False, 'result': 'clean site'}, 'Malware Domain Blocklist': {'detected': False, 'result': 'clean site'}, 'MalwarePatrol': {'detected': False, 'result': 'clean site'}, 'Webu
tation': {'detected': False, 'result': 'clean site'}, 'Trustwave': {'detected': False, 'result': 'clean site'}, 'Web Security Guard': {'detected': False, 'result': 'clean site'}, 'CyRadar'
: {'detected': False, 'result': 'clean site'}, 'desenmascara.me': {'detected': False, 'result': 'clean site'}, 'ADMINUSLabs': {'detected': False, 'result': 'clean site'}, 'Malwarebytes hpH
osts': {'detected': False, 'result': 'clean site'}, 'Dr.Web': {'detected': False, 'result': 'clean site'}, 'AlienVault': {'detected': False, 'result': 'clean site'}, 'Emsisoft': {'detected
': False, 'result': 'clean site'}, 'Rising': {'detected': False, 'result': 'clean site'}, 'Malc0de Database': {'detected': False, 'result': 'clean site', 'detail': 'http://malc0de.com/data
base/index.php?search=alicdn.re.kr'}, 'malwares.com URL checker': {'detected': False, 'result': 'clean site'}, 'Phishtank': {'detected': False, 'result': 'clean site'}, 'Malwared': {'detec
ted': False, 'result': 'clean site'}, 'Avira': {'detected': False, 'result': 'clean site'}, 'StopBadware': {'detected': False, 'result': 'unrated site'}, 'Antiy-AVL': {'detected': False, '
result': 'clean site'}, 'Forcepoint ThreatSeeker': {'detected': False, 'result': 'clean site'}, 'SCUMWARE.org': {'detected': False, 'result': 'clean site'}, 'Comodo Site Inspector': {'dete
cted': False, 'result': 'clean site'}, 'Malekal': {'detected': False, 'result': 'clean site'}, 'ESET': {'detected': False, 'result': 'clean site'}, 'Sophos': {'detected': False, 'result':
'unrated site'}, 'Yandex Safebrowsing': {'detected': False, 'result': 'clean site', 'detail': 'http://yandex.com/infected?l10n=en&url=http://alicdn.re.kr/re.kr'}, 'Spam404': {'detected': F
alse, 'result': 'clean site'}, 'Nucleon': {'detected': False, 'result': 'clean site'}, 'Sucuri SiteCheck': {'detected': True, 'result': 'malicious site'}, 'Blueliv': {'detected': False, 'r
esult': 'clean site'}, 'Netcraft': {'detected': False, 'result': 'unrated site'}, 'AutoShun': {'detected': False, 'result': 'unrated site'}, 'ThreatHive': {'detected': False, 'result': 'cl
ean site'}, 'FraudScore': {'detected': False, 'result': 'clean site'}, 'Tencent': {'detected': False, 'result': 'clean site'}, 'URLQuery': {'detected': False, 'result': 'unrated site'}, 'F
ortinet': {'detected': False, 'result': 'clean site'}, 'ZeroCERT': {'detected': False, 'result': 'clean site'}, 'Baidu-International': {'detected': False, 'result': 'clean site'}, 'securol
ytics': {'detected': False, 'result': 'clean site'}}}
```

SAMZO

[바이러스토탈]

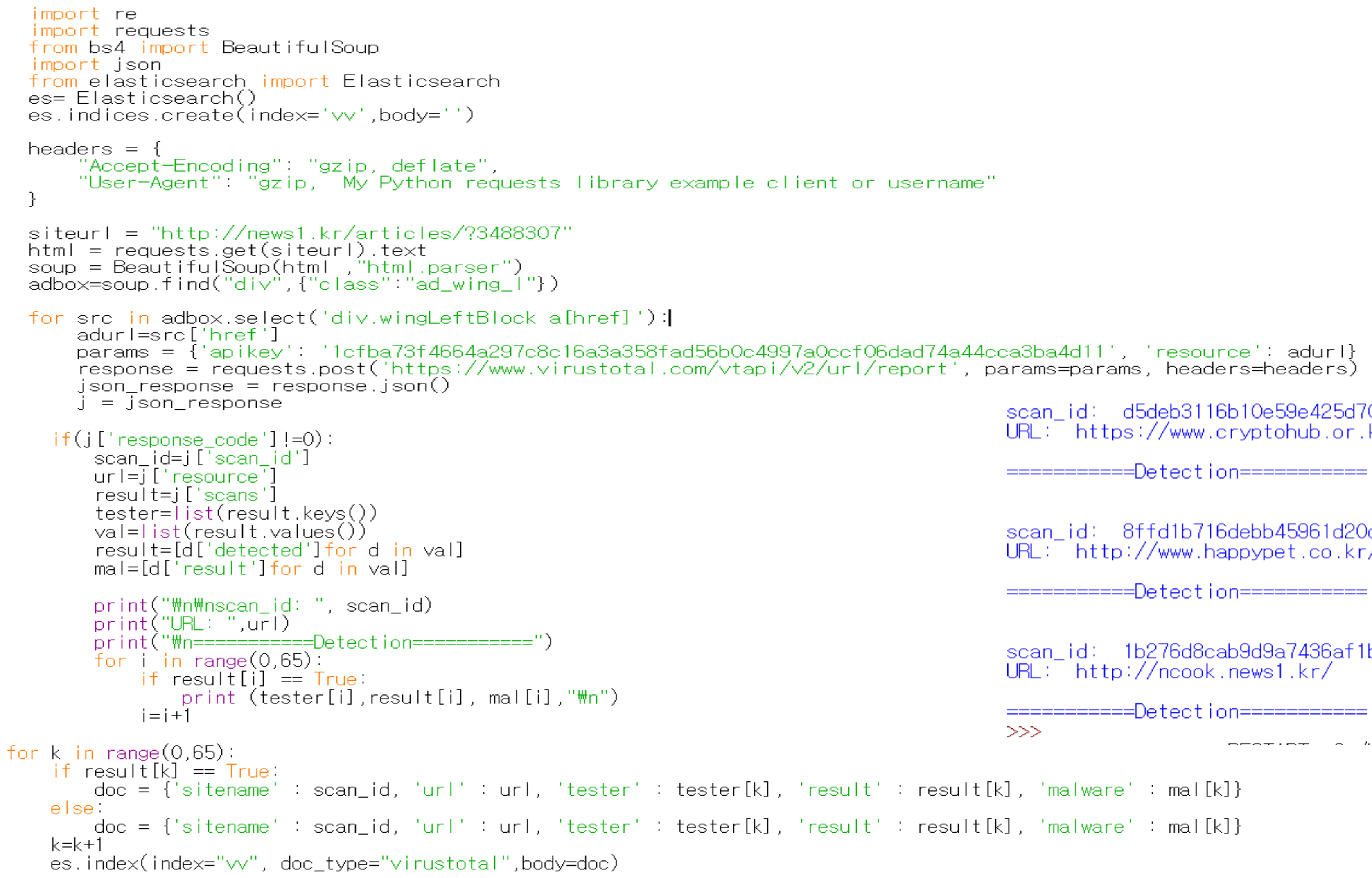
바이러스토탈 분석에서 malware가 뜨는 결과만 추출

```python
import requests
headers = {
    "Accept-Encoding": "gzip, deflate",
    "User-Agent": "gzip,  My Python requests library example client or username"
}
params = {'apikey': '1cfba73f4664a297c8c16a3a358fad56b0c4997a0ccf06dad74a44cca3ba4d11', 'resource':'http://alicdn.re.kr/re.kr'}
response = requests.post('https://www.virustotal.com/vtapi/v2/url/report', params=params, headers=headers)
json_response = response.json()
j = json_response

if(j['response_code']!=0):
        scan_id=j['scan_id']
        url=j['resource']
        result=j['scans']
        tester=list(result.keys())
        val=list(result.values())
        result=[d['detected']for d in val]
        mal=[d['result']for d in val]

        print("₩n₩nscan_id: ", scan_id)
        print("URL: ",url)
        print("₩n===========Detection===========")
        for i in range(0,65):
            if result[i] == True:
                print (tester[i],result[i], mal[i],"₩n")

            i=i+1
```

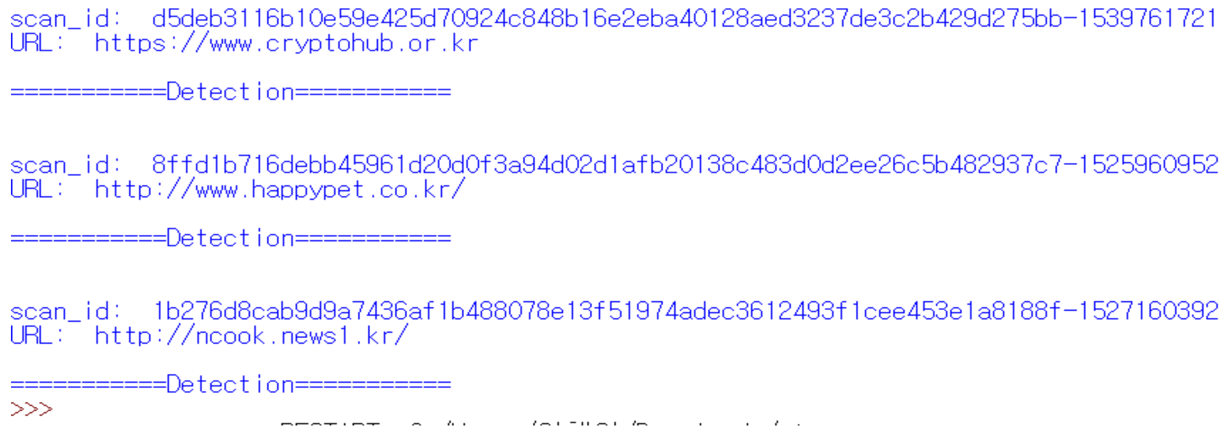
```
scan_id:  f9f6233c1840eb98c87f7df7fc85927dd4cc0f31faac32b7a960ef5cc8c66f61-1514263742
URL:  http://alicdn.re.kr/re.kr

===========Detection===========
Sucuri SiteCheck True malicious site
```

SAMZO

[바이러스토탈]



```
import re
import requests
from bs4 import BeautifulSoup
import json
from elasticsearch import Elasticsearch
es= Elasticsearch()
es.indices.create(index='vv',body='')

headers = {
    "Accept-Encoding": "gzip, deflate",
    "User-Agent" : "gzip,  My Python requests library example client or username"
}

siteurl = "http://news1.kr/articles/?3488307"
html = requests.get(siteurl).text
soup = BeautifulSoup(html ,"html.parser")
adbox=soup.find("div",{"class":"ad_wing_l"})

for src in adbox.select('div.wingLeftBlock a[href]'):
    adurl=src['href']
    params = {'apikey': '1cfba73f4664a297c8c16a3a358fad56b0c4997a0ccf06dad74a44cca3ba4d11', 'resource': adurl}
    response = requests.post('https://www.virustotal.com/vtapi/v2/url/report', params=params, headers=headers)
    json_response = response.json()
    j = json_response

  if(j['response_code']!=0):
      scan_id=j['scan_id']
      url=j['resource']
      result=j['scans']
      tester=list(result.keys())
      val=list(result.values())
      result=[d['detected']for d in val]
      mal=[d['result']for d in val]

      print("₩n₩nscan_id: ", scan_id)
      print("URL: ",url)
      print("₩n===========Detection==========")
      for i in range(0,65):
          if result[i] == True:
              print (tester[i],result[i], mal[i],"₩n")
          i=i+1

for k in range(0,65):
    if result[k] == True:
        doc = {'sitename' : scan_id, 'url' : url, 'tester' : tester[k], 'result' : result[k], 'malware' : mal[k]}
    else:
        doc = {'sitename' : scan_id, 'url' : url, 'tester' : tester[k], 'result' : result[k], 'malware' : mal[k]}
    k=k+1
    es.index(index="vv", doc_type="virustotal",body=doc)
```

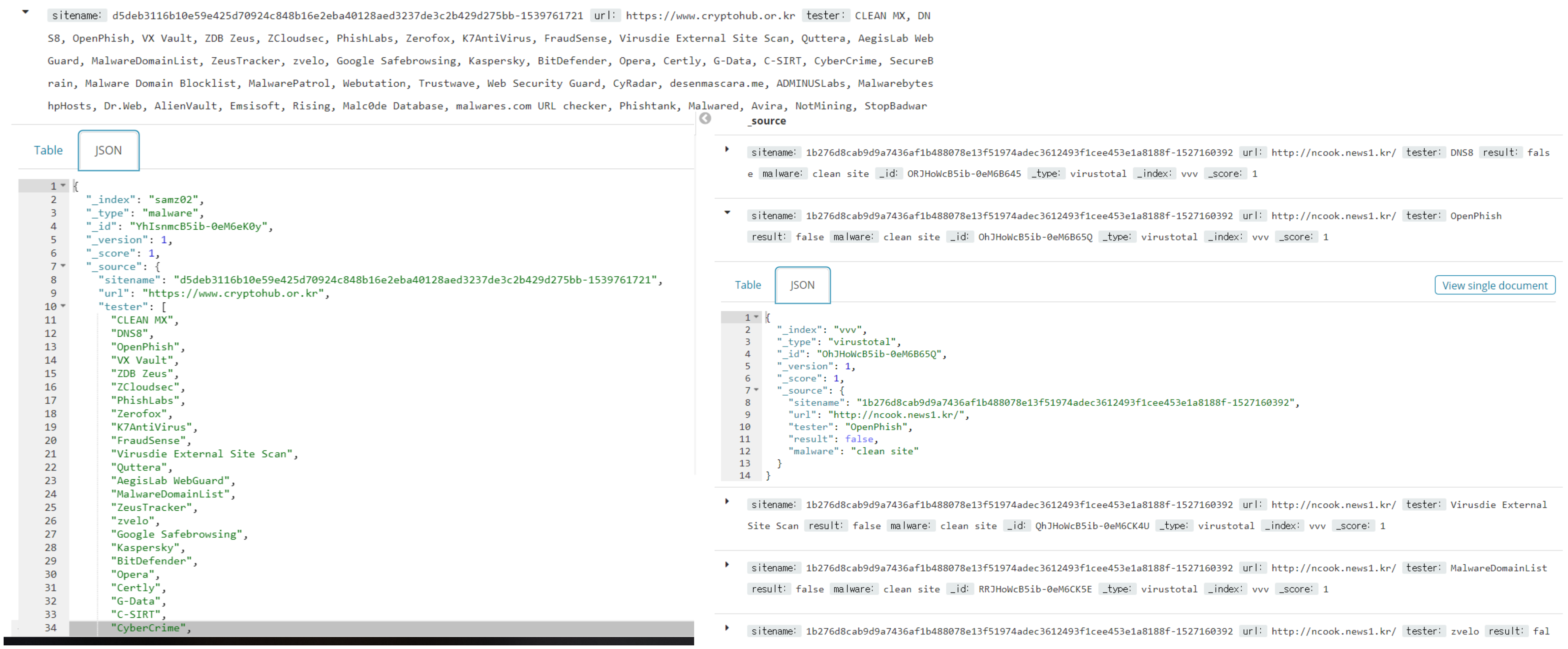


```
scan_id:  d5deb3116b10e59e425d70924c848b16e2eba40128aed3237de3c2b429d275bb-1539761721
URL:  https://www.cryptohub.or.kr

===========Detection==========


scan_id:  8ffd1b716debb45961d20d0f3a94d02d1afb20138c483d0d2ee26c5b482937c7-1525960952
URL:  http://www.happypet.co.kr/

===========Detection==========


scan_id:  1b276d8cab9d9a7436af1b488078e13f51974adec3612493f1cee453e1a8188f-1527160392
URL:  http://ncook.news1.kr/

===========Detection==========
>>>
```

SAMZO

[바이러스토탈]

Elasticsearch DB에 삽입한 데이터를 kibana로 연동시킴



```
sitename: d5deb3116b10e59e425d70924c848b16e2eba40128aed3237de3c2b429d275bb-1539761721 url: https://www.cryptohub.or.kr tester: CLEAN MX, DN
S8, OpenPhish, VX Vault, ZDB Zeus, ZCloudsec, PhishLabs, Zerofox, K7AntiVirus, FraudSense, Virusdie External Site Scan, Quttera, AegisLab Web
Guard, MalwareDomainList, ZeusTracker, zvelo, Google Safebrowsing, Kaspersky, BitDefender, Opera, Certly, G-Data, C-SIRT, CyberCrime, SecureB
rain, Malware Domain Blocklist, MalwarePatrol, Webutation, Trustwave, Web Security Guard, CyRadar, desenmascara.me, ADMINUSLabs, Malwarebytes
hpHosts, Dr.Web, AlienVault, Emsisoft, Rising, Malc0de Database, malwares.com URL checker, Phishtank, Malwared, Avira, NotMining, StopBadwar
```

Table JSON

```
{
  "_index": "samz02",
  "_type": "malware",
  "_id": "YhIsnmcB5ib-0eM6K0y",
  "_version": 1,
  "_score": 1,
  "_source": {
    "sitename": "d5deb3116b10e59e425d70924c848b16e2eba40128aed3237de3c2b429d275bb-1539761721",
    "url": "https://www.cryptohub.or.kr",
    "tester": [
      "CLEAN MX",
      "DNS8",
      "OpenPhish",
      "VX Vault",
      "ZDB Zeus",
      "ZCloudsec",
      "PhishLabs",
      "Zerofox",
      "K7AntiVirus",
      "FraudSense",
      "Virusdie External Site Scan",
      "Quttera",
      "AegisLab WebGuard",
      "MalwareDomainList",
      "ZeusTracker",
      "zvelo",
      "Google Safebrowsing",
      "Kaspersky",
      "BitDefender",
      "Opera",
      "Certly",
      "G-Data",
      "C-SIRT",
      "CyberCrime",
```

_source

```
sitename: 1b276d8cab9d9a7436af1b488078e13f51974adec3612493f1cee453e1a8188f-1527160392 url: http://ncook.news1.kr/ tester: DNS8 result: fals
e malware: clean site _id: ORJHoWcB5ib-0eM6B645 _type: virustotal _index: vvv _score: 1

sitename: 1b276d8cab9d9a7436af1b488078e13f51974adec3612493f1cee453e1a8188f-1527160392 url: http://ncook.news1.kr/ tester: OpenPhish
result: false malware: clean site _id: OhJHoWcB5ib-0eM6B65Q _type: virustotal _index: vvv _score: 1
```

Table JSON View single document

```
{
  "_index": "vvv",
  "_type": "virustotal",
  "_id": "OhJHoWcB5ib-0eM6B65Q",
  "_version": 1,
  "_score": 1,
  "_source": {
    "sitename": "1b276d8cab9d9a7436af1b488078e13f51974adec3612493f1cee453e1a8188f-1527160392",
    "url": "http://ncook.news1.kr/",
    "tester": "OpenPhish",
    "result": false,
    "malware": "clean site"
  }
}
```

```
sitename: 1b276d8cab9d9a7436af1b488078e13f51974adec3612493f1cee453e1a8188f-1527160392 url: http://ncook.news1.kr/ tester: Virusdie External
Site Scan result: false malware: clean site _id: QhJHoWcB5ib-0eM6CK4U _type: virustotal _index: vvv _score: 1

sitename: 1b276d8cab9d9a7436af1b488078e13f51974adec3612493f1cee453e1a8188f-1527160392 url: http://ncook.news1.kr/ tester: MalwareDomainList
result: false malware: clean site _id: RRJHoWcB5ib-0eM6CK5E _type: virustotal _index: vvv _score: 1

sitename: 1b276d8cab9d9a7436af1b488078e13f51974adec3612493f1cee453e1a8188f-1527160392 url: http://ncook.news1.kr/ tester: zvelo result: fal
```

SAMZO

3. 한계점&어려웠던 점

SAMZO

[한계점]

✓ 사이트마다 html 태그나 클래스명, 구조 등 달라서 일반화할 수가 없었다.
- 다른 사이트를 돌릴 때 마다 코드의 수정이 필요했고, 크롤링할 때 마다 힘들게 찾아야했다.

✓ 실제 사이트 광고에는 악성코드가 발견되지 않아 실질적으로 구현하고자 하는 결과를 확인할 수가 없다.
- 악성URL만 따로 바이러스토탈에 돌려 실제 악성코드 발견시에는 어떻게 결과가 나오는지 부분적으로 확인이 가능하다.

SAMZO

[어려웠던 점]

✓ 윈도우 환경에서의 활용이 어디에도 나와있지 않아 창조적으로 시도하게 되었다.

✓ 엘라스틱서치에 관련된 내용들이 리눅스기반 사용이어 서 윈도우 버전에 맞게 새로 코드를 작성해야 했다.

✓ 바이러스토탈 결과값에서 필요한 정보를 뽑기 위해 어떤 형식으로 출력되는지 실제로 코드를 짜면서 확인했어야 했다.

SAMZO

[참고자료]

+ Google Search
+ 교수님

SAMZO

감사합니다