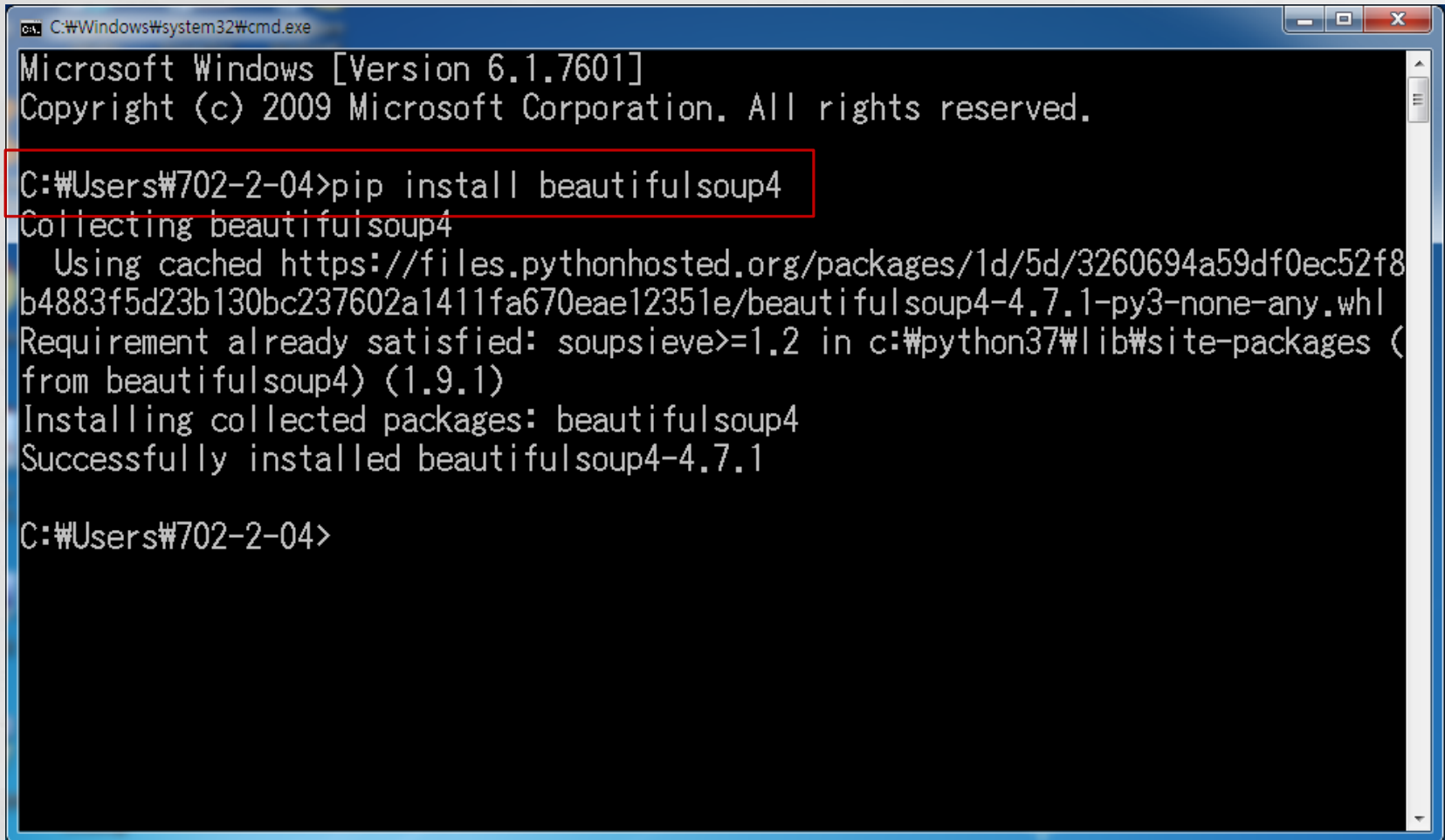


# Chap08. Web Crawling

작성자 : 김진성

# 패키지 설치

pip install beautifulsoup4

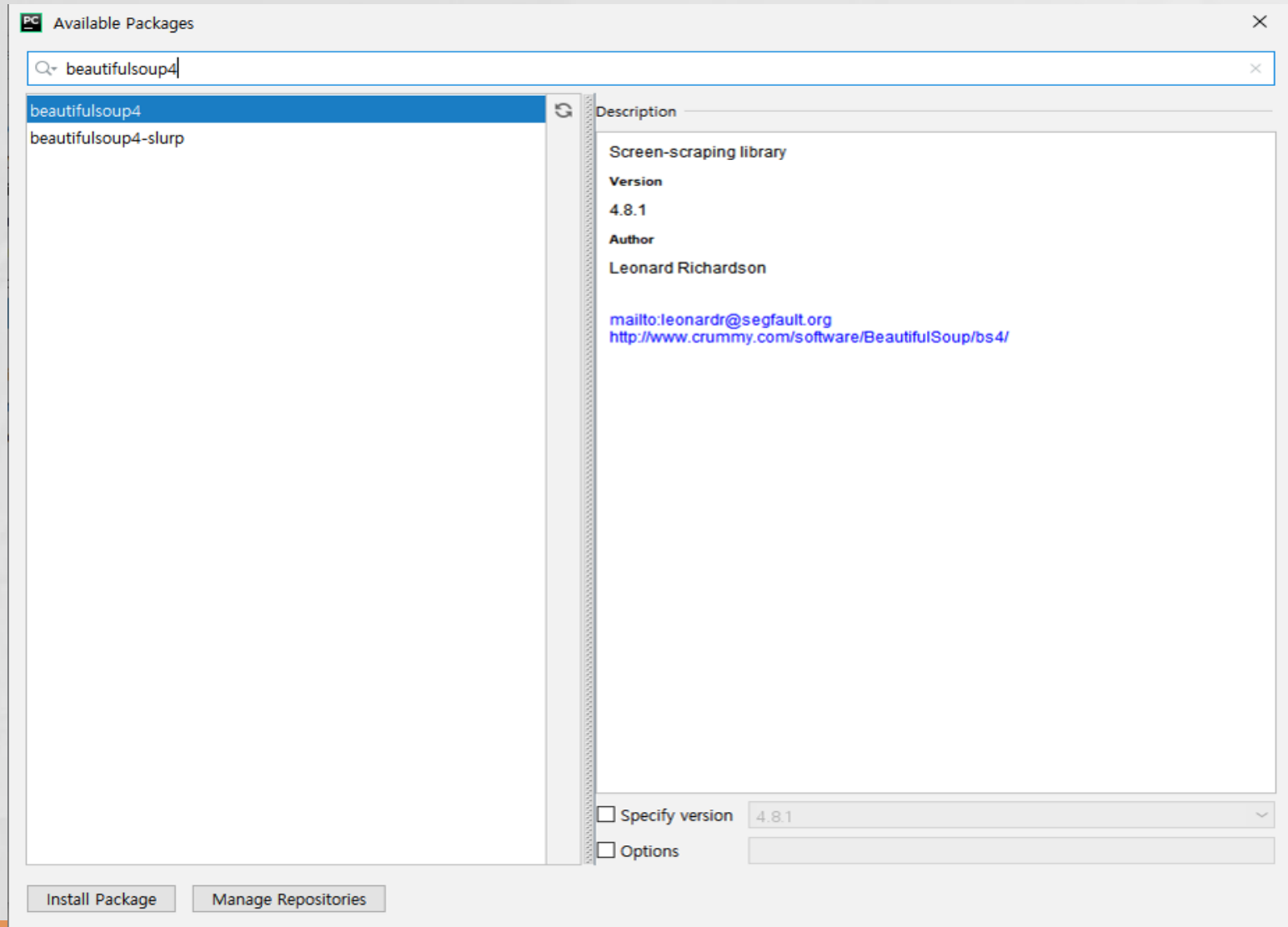


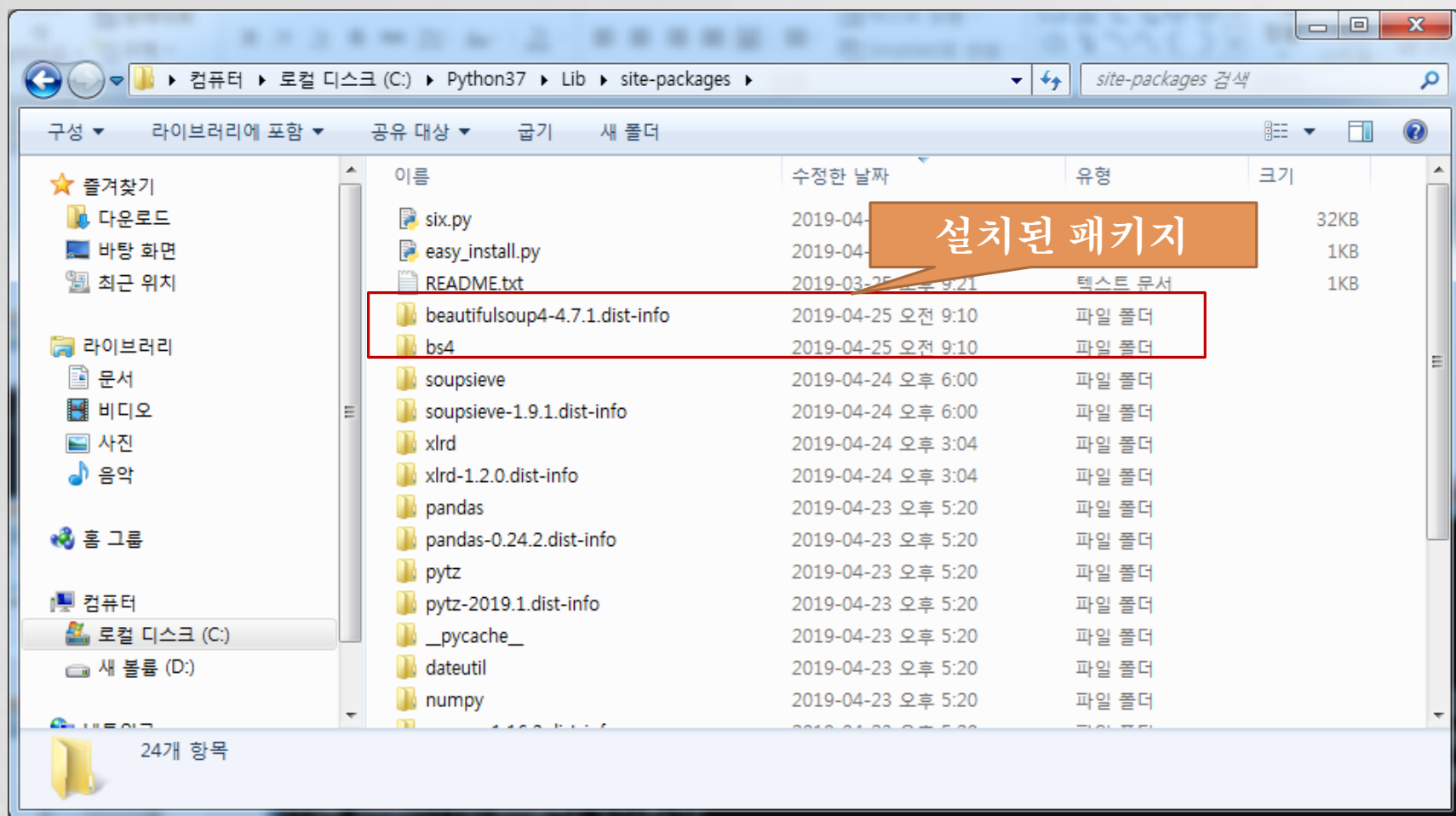
```
C:\Windows\system32\cmd.exe
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\702-2-04>pip install beautifulsoup4
Collecting beautifulsoup4
  Using cached https://files.pythonhosted.org/packages/1d/5d/3260694a59df0ec52f8b4883f5d23b130bc237602a1411fa670eae12351e/beautifulsoup4-4.7.1-py3-none-any.whl
Requirement already satisfied: soupsieve>=1.2 in c:\python37\lib\site-packages (from beautifulsoup4) (1.9.1)
Installing collected packages: beautifulsoup4
Successfully installed beautifulsoup4-4.7.1

C:\Users\702-2-04>
```

# Pycharm 설치





# Html Parsing Web Crawling

```
import urllib.request # url 요청 모듈
from lxml.html import parse # html 양식으로 파싱
from io import StringIO # 문자열 입출력 모듈
```

```
# 1. web 문서를 source(text문서) 로 가져오기
url = "http://media.daum.net/"
#url = "http://news.naver.com/"
```

```
# 1) html source 가져오기
res = urllib.request.urlopen(url) # web 문서 get
# requests.get(url)
data = res.read() # binary 형태로 읽음
#print(data) # b'\n<!doctype html>\n'
```

```
# 2) html 문서열로 변환(파싱)
text = data.decode("utf-8")
text_source = StringIO(text)
parsed = parse(text_source)
print(parsed)
```

```
# 3) root node 찾기
root_node = parsed.getroot()
```



# 2. html의 <a>태그 가져오기

# 형식) root\_node.findall("./태그")

links = root\_node.findall("./a")

print('링크수: ', len(links)) # 링크수: 202

print(links) # 202 링크 element object

# 3. 'href' 속성값 가져오기

# 형식) obj.get('속성')

link\_url = [] # 속성값을 저장

cnt = 1

for link in links :

print(cnt, '->', link.get('href'))

link\_url.append(link.get('href')) # 내용 추가

cnt += 1

print(link\_url) # 전체 내용 출력

# 4. <a>태그 내용 가져오기

cnt = 1

centents = []

for link in links :

print(cnt, '->', link.text\_content().strip())

cnt += 1

centents.append(link.text\_content().strip())

# BeautifulSoup Web Crawling

```
import urllib.request
from bs4 import BeautifulSoup
```

```
url = 'http://localhost:8282/DataCrawlingServer/html/html01.html'
```

# 1. html source 가져오기

```
res = urllib.request.urlopen(url) # web 문서 get
data = res.read() # binary 형태로 읽음
```

# 2. html 파싱

```
html = data.decode("utf-8") # 디코딩
soup = BeautifulSoup(html, 'html.parser') # html source 파싱
```

# 3. 태그 내용 가져오기

# 1) 태그 <h1> 가져오기

```
h1 = soup.html.body.h1
print('h1 :', h1.string) # h1 : 시멘틱 태그?
```

# 2) find() 함수로 찾기

```
h2 = soup.find("h2")
print("h2 :", h2.string) # h2 : 주요 시멘틱 태그
```

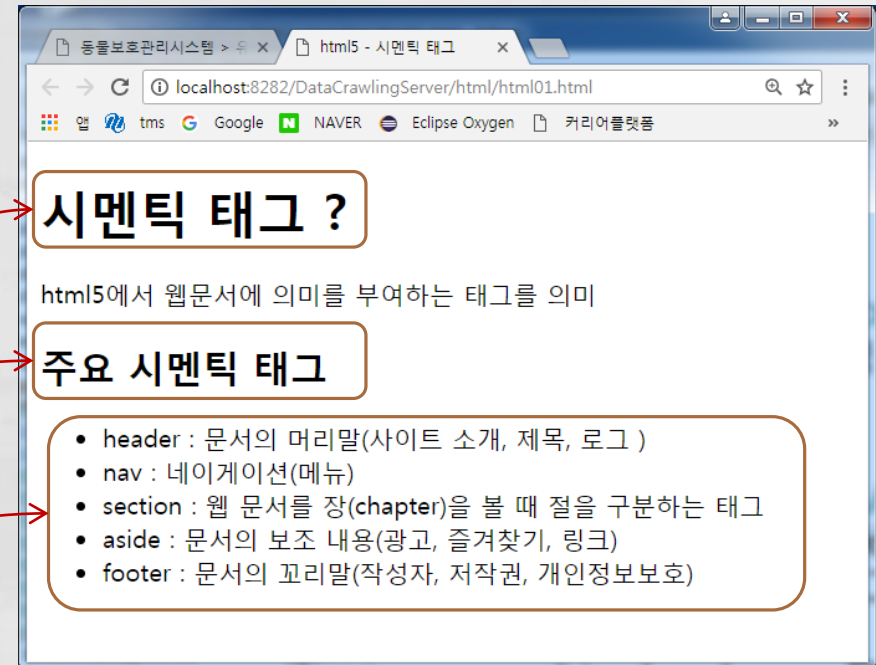
```
li = soup.find("li")
```

```
print(li.string) # header : 문서의 머리말(사이트 소개, 제목, 로그)
```

# 2) find\_all() 함수로 여러개 찾기 : list 반환

```
li2 = soup.find_all("li")
print(li2) # [<li> header : 문서의 머리말(사이트 소개, 제목, 로그)</li>, ...]
# print(li2.string) # error 발생
```

```
for li in li2 :
    print(li.string)
```



# 유기동물.동물보호센터

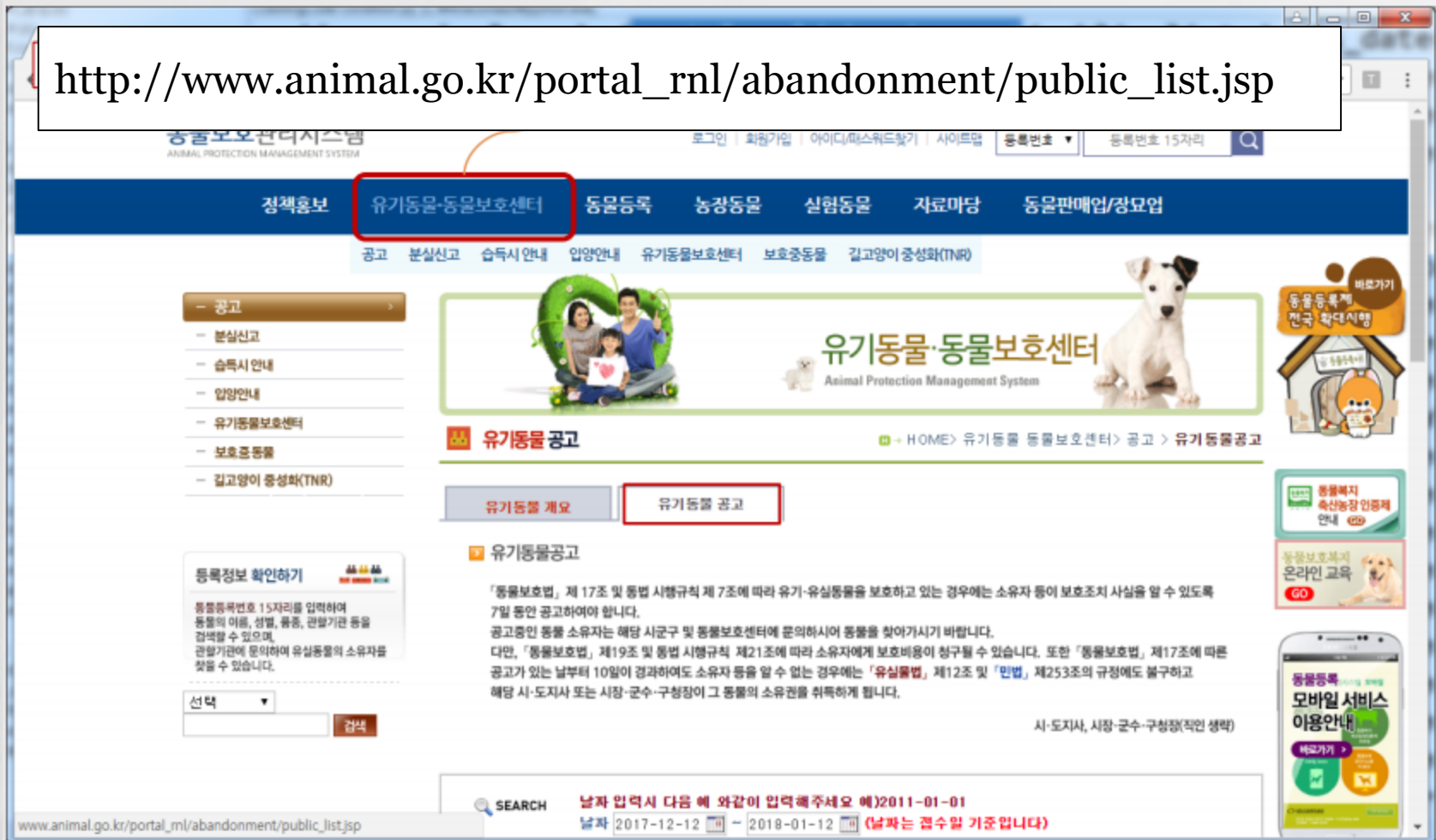
http://www.animal.go.kr/portal\_rnl/index.jsp

The screenshot displays the homepage of the Animal Protection Center. At the top, there is a browser address bar showing the URL [http://www.animal.go.kr/portal\\_rnl/index.jsp](http://www.animal.go.kr/portal_rnl/index.jsp). Below the browser window, the website header includes the logo "동물보호관리시스템" and a navigation menu with links: 정책홍보, 유기동물·동물보호센터, 동물등록, 농경동물, 상형동물, 자원매당, 동물단체입/양요업. The main banner features a photograph of a young girl interacting with a dog, with the text "소중한 생명! 여러분의 반려동물 사랑으로 지켜주세요" (Precious life! Please protect your pet with your love). Below the banner, there are several service sections: "반려동물물려주고 싶으세요?" (Do you want to give your pet?), "유기동물 공고" (Lost pet notice), "찾아주세요" (Find it), and "동물보호센터 검색" (Search for animal protection centers). The bottom of the page contains more detailed information about animal protection services, including a section for "동물등록제 전국 확대" (National expansion of the pet registration system).



# [유기동물공고] 페이지

[http://www.animal.go.kr/portal\\_rnl/abandonment/public\\_list.jsp](http://www.animal.go.kr/portal_rnl/abandonment/public_list.jsp)



## 검색조건 : 날짜/시군구/축종/상태 검색

동물보호관리시스템 > X

www.animal.go.kr/portal\_m/abandonment/public\_list.jsp

SEARCH

날짜 입력시 다음 해 외같이 입력해주세요 예) 2011-01-01  
 날짜 2017-12-12 ~ 2018-01-12 (날짜는 콤수일 기준입니다)  
 시도 전체 시군구 선택 보호센터 전체  
 축종 전체 선택 상태 전체 조회

※ 검색시 유의사항 : 동종오류가 발생할 수 있으니 축종을 전체로 설정 후 한번 더 검색하시기 바랍니다.  
 ※ 광고종인 동물 소유자는 "자세히 보기"를 참고하시어 해당 시군구 및 동물보호센터 또는 동물보호상담센터 1577-0954 로 문의하시기 바랍니다.  
 ※ 동물보호센터 및 동물병원 근무시간은 09:00 ~ 18:00이므로 문의전화는 근무시간에만 가능합니다.

전체 조회건수 : 6716(건)

 <p>자세히 보기</p>	<p>공고번호 서울-서울-2018-00019</p> <p>접수일 2018-01-12</p> <p>품종 불명</p> <p>성별 수컷</p> <p>발견장소 강변면 대학교</p> <p>특징 입양희한 목줄</p> <p>상태 미검역</p>	 <p>자세히 보기</p>	<p>공고번호 경남-남해-2018-00005</p> <p>접수일 2018-01-12</p> <p>품종 믹스견</p> <p>성별 암컷</p> <p>발견장소 남해군 남해읍 전소..</p> <p>특징 관순함, 검게심이 ..</p> <p>상태 공고중</p>
 <p>자세히 보기</p>	<p>공고번호 경남-고성-2018-00011</p> <p>접수일 2018-01-12</p> <p>품종 믹스견</p> <p>성별 암컷</p> <p>발견장소 경남 고성군 통해..</p> <p>특징 암초 수2</p> <p>상태 미검역</p>	 <p>자세히 보기</p>	<p>공고번호 경남-사천-2018-00019</p> <p>접수일 2018-01-12</p> <p>품종 믹스견</p> <p>성별 수컷</p> <p>발견장소 사천시 한주아파트</p> <p>특징 전좌 2개월 추정</p> <p>상태 공고중</p>
 <p>자세히 보기</p>	<p>공고번호 경남-사천-2018-00018</p> <p>접수일 2018-01-12</p> <p>품종 믹스견</p> <p>성별 수컷</p> <p>발견장소 사천시 진삼로 12..</p> <p>특징 왼쪽 견강</p> <p>상태 미검역</p>	 <p>자세히 보기</p>	<p>공고번호 경북-성주-2018-00010</p> <p>접수일 2018-01-12</p> <p>품종 믹스견</p> <p>성별 이상</p> <p>발견장소 신원소방서 통보</p> <p>특징 황색 믹스견 강아지..</p> <p>상태 미검역</p>
 <p>자세히 보기</p>	<p>공고번호 전남-순천-2018-00023</p>	 <p>자세히 보기</p>	<p>공고번호 전남-순천-2018-00022</p>

동물등록 전국 확대 시행

동물복지 축산농장 인증제 안내

동물보호법 온라인 교육

동물등록 모바일 서비스 이용안내

# 검색조건 : 2015~2018년도/서울시/강남구/개

SEARCH

날짜 입력시 다음 예와 같이 입력해주세요 예)2011-01-01

날짜 2015-01-01 ~ 2018-01-12 (날짜는 필수일 기준입니다)

시도 서울특별시 시군구 강남구 보호센터

전체

속종 개 선택 상태 전체 조회

- ※ 검색시 유의사항 : 품종유가 발생할 수 있으니 속종을 전체로 설정 후 한번 더 검색하시기 바랍니다.  
 ※ 공고중인 동물 소유자는 "자세히 보기"를 참고하시어 해당 시군구 및 동물보호센터 또는 동물보호상담센터 1577-0954 로 문의하시기 바랍니다.  
 ※ 동물보호센터 및 동물병원 근무시간은 09:00 - 18:00이므로 문의전화는 근무시간에만 가능합니다.

전체 조회건수 : 475(건)



자세히 보기

공고번호 서울-강남-2018-00008  
 접수일 2018-01-11  
 품종 푸들  
 성별 수컷  
 발견장소 노원1동 인근  
 특징 양귀/얼굴털남기고전..  
 상태 종료(반환)



자세히 보기

공고번호 서울-강남-2018-00007  
 접수일 2018-01-09  
 품종 믹스견  
 성별 암컷  
 발견장소 삼성동 삼성중앙역..  
 특징 양귀처럼. 코검정..  
 상태 공고중



자세히 보기

공고번호 서울-강남-2018-00006  
 접수일 2018-01-09  
 품종 텍스폰트  
 성별 암컷  
 발견장소 강남구청  
 특징 장모종. 코갈색.유선..  
 상태 공고중



자세히 보기

공고번호 서울-강남-2018-00004  
 접수일 2018-01-03  
 품종 시츄  
 성별 암컷  
 발견장소 역삼동 차도  
 특징 고형.전신파부질판..  
 상태 종료(자연사)



자세히 보기

공고번호 서울-강남-2018-00003  
 접수일 2018-01-03  
 품종 푸들  
 성별 수컷  
 발견장소 도곡동 416-7..  
 특징 백내강.코갈색.전신..  
 상태 공고중



자세히 보기

공고번호 서울-강남-2018-00002  
 접수일 2018-01-03  
 품종 보스턴 테리어  
 성별 암컷  
 발견장소 역삼동 경복아파트..  
 특징 코검정.피부각질.사..  
 상태 종료(반환)



자세히 보기

공고번호 서울-강남-2018-00001  
 접수일 2017-12-30  
 품종 푸들  
 성별 수컷  
 발견장소 노원동 176-4..  
 특징 노랑자국.배검황갈..  
 상태 공고중



자세히 보기

공고번호 서울-강남-2017-00243  
 접수일 2017-12-18  
 품종 텍스폰트  
 성별 암컷  
 발견장소 개포동 12-2  
 특징 양귀다리뺏음.원뿔다..  
 상태 공고중





http://www.animal.go.kr/portal\_rnl/abandonment/public\_list.jsp?s\_date=2015-01-01&e\_date=2018-01-12  
&s\_upr\_cd=6110000&s\_org\_cd=3220000&s\_up\_kind\_cd=417000&s\_kind\_cd=&s\_name=&s\_shelter\_cd=&s\_wrk\_cd=&s\_state=&s\_state\_hidden=&pagecnt=48

조건검색에 따른 URL  
검색년도 : s\_date&e\_date  
검색시도 : s\_upr\_cd=6110000  
검색 시군구 :s\_org\_cd=3220000  
검색페이지 : pagecnt=48

SEARCH 날짜 입력시 다음 예와같이 입력해주세요 예)2011-01-01  
날짜 2015-01-01 ~ 2018-01-12 (날짜는 필수일 기준입니다)  
시도 서울특별시 시군구 강남구 보호센터  
전체  
속종 개 선택 상태 전체 조회

※ 검색시 유의사항 : 품종오류가 발생할 수 있으니 속종을 전체로 설정 후 한번 더 검색하시기 바랍니다.  
※ 공고중인 동물 소유자는 "자세히 보기"를 참고하시어 해당 시군구 및 동물보호센터 또는 동물보호상담센터 1577-0954 로 문의하시기 바랍니다.  
※ 동물보호센터 및 동물병원 근무시간은 09:00 ~ 18:00이므로 문의전화는 근무시간에만 가능합니다.

> 전체 조회건수 :475(건)



공고번호 서울-강남-2015-00007  
접수일 2015-01-14  
품종 푸들  
성별 수컷  
발견장소 역삼동 798-20.  
특징 얼굴탈팔음,코연한팔..  
상태 종료(반환)

자세히 보기



공고번호 서울-강남-2015-00006  
접수일 2015-01-12  
품종 기타  
성별 암컷  
발견장소 매치4동 성당 인근..  
특징 빨간바탕에양옆에검정..  
상태 종료(반환)

자세히 보기



공고번호 서울-강남-2015-00005  
접수일 2015-01-11  
품종 푸들  
성별 수컷  
발견장소 역삼동 705-25.  
특징 눈 주변탈팔음,코검정..  
상태 종료(반환)

자세히 보기



공고번호 서울-강남-2015-00004  
접수일 2015-01-11  
품종 푸들  
성별 수컷  
발견장소 역삼역 1번출구 인근..  
특징 설사,좌후지발바닥상..  
상태 종료(입양)

자세히 보기



공고번호 서울-강남-2015-00003  
접수일 2015-01-03  
품종 믹스견  
성별 수컷  
발견장소 수서경찰서 인근  
특징 빨간바탕에노란줄2개..  
상태 종료(반환)

자세히 보기

40 41 42 43 44 45 46 47 48 49

전체 검색 페이지 48 페이지[현재 : 48page]



동물보호관리시스템 > X

www.animal.go.kr/portal\_rn/abandonment/public\_list.jsp?s\_date=2015-01-01&e\_date=2018-01-12&s\_upr\_cd=6110000&s\_org\_cd=0000000&s\_up\_kind\_cd=&s\_kind\_cd=&s\_name=&s\_shelter...

7월 동안 공고하여야 합니다.  
공고중인 동물 소유자는 해당 시군구 및 동물보호센터에 문의하시어 동물을 찾아가시기 바랍니다.

## 검색조건 : 2015~2018년도/서울시/전체/전체

사·도지사, 시장·군수·구청장직인 생애

전국 확대사항

SEARCH 날짜 2015-01-01 ~ 2018-01-12 (날짜는 필수입력 기준입니다)  
시도 서울특별시 시군구 선택 보호센터 전체  
속종 전체 선택 상태 전체 조회

※ 검색시 유의사항 : 품종오류가 발생할 수 있으니 속종을 전체로 설정 후 한번 더 검색하시기 바랍니다.  
※ 공고중인 동물 소유자는 "자세히 보기"를 참고하시어 해당 시군구 및 동물보호센터 또는 동물보호상담센터 1577-0954 로 문의하시기 바랍니다.  
※ 동물보호센터 및 동물병원 근무시간은 09:00 ~ 18:00이므로 문의전화는 근무시간에만 가능합니다.

전체 조회건 수 : 26273(건)

공고번호 서울-양천-2015-00004  
접수일 2015-01-01  
품종 시츄  
성별 암컷  
발견장소 신정7동봉영중학교앞...  
특징 치석있고부절 교합이며...  
상태 종료(반환)

자세히 보기

공고번호 서울-종산-2015-00003  
접수일 2015-01-01  
품종 고양이  
성별 수컷  
발견장소 종산구 소월로 40...  
특징 후지 마비  
상태 종료(자연사)

자세히 보기

공고번호 서울-종산-2015-00002  
접수일 2015-01-01  
품종 알라트  
성별 수컷  
발견장소 이촌아파트 중간 도...  
특징 눈물입고있음  
상태 종료(반환)

자세히 보기

2621 2622 2623 2624 2625 2626 2627 2628

이용안내 | 개인정보처리방침 | 저작권 정책  
(우)39660 경상북도 김천시 학신8로 177(출국동) 업무문의: 054-912-0518, 동물보호상담센터: 1577-0954 | loveanimal@korea.kr  
copyright by Animal and Plant Quarantine Agency. All Rights Reserved.

농림축산검역본부  
Animal and Plant Quarantine Agency

동물보호  
WA  
WEB ACCESSIBILITY

[http://www.animal.go.kr/portal\\_rnl/abandonment/public\\_list.jsp?s\\_date=2015-01-01&e\\_date=2018-01-12&s\\_upr\\_cd=6110000&s\\_org\\_cd=0000000&s\\_up\\_kind\\_cd=&s\\_kind\\_cd=&s\\_name=&s\\_shelter\\_cd=&s\\_wrk\\_cd=&s\\_state=&s\\_state\\_hidden=&pagecnt=2628](http://www.animal.go.kr/portal_rnl/abandonment/public_list.jsp?s_date=2015-01-01&e_date=2018-01-12&s_upr_cd=6110000&s_org_cd=0000000&s_up_kind_cd=&s_kind_cd=&s_name=&s_shelter_cd=&s_wrk_cd=&s_state=&s_state_hidden=&pagecnt=2628)

조건검색에 따른 URL

검색년도 : s\_date&e\_date

검색시도 : s\_upr\_cd=6110000

검색 시군구 : s\_org\_cd=0000000

검색페이지 : pagecnt=2628

서울시 전체 페이지 : 2628 page



이용안내 | 개인정보처리방침 | 저작권정책

(우)139660 경상북도 김천시 월신8로 177(율곡동) 업무문의: 054-912-0318, 동물보호상담센터: 1577-0954, [loveanimal.go.kr](http://loveanimal.go.kr)  
copyright by Animal and Plant Quarantine Agency. All Rights Reserved.





# 유기견 자료 Crawling 대상 문서

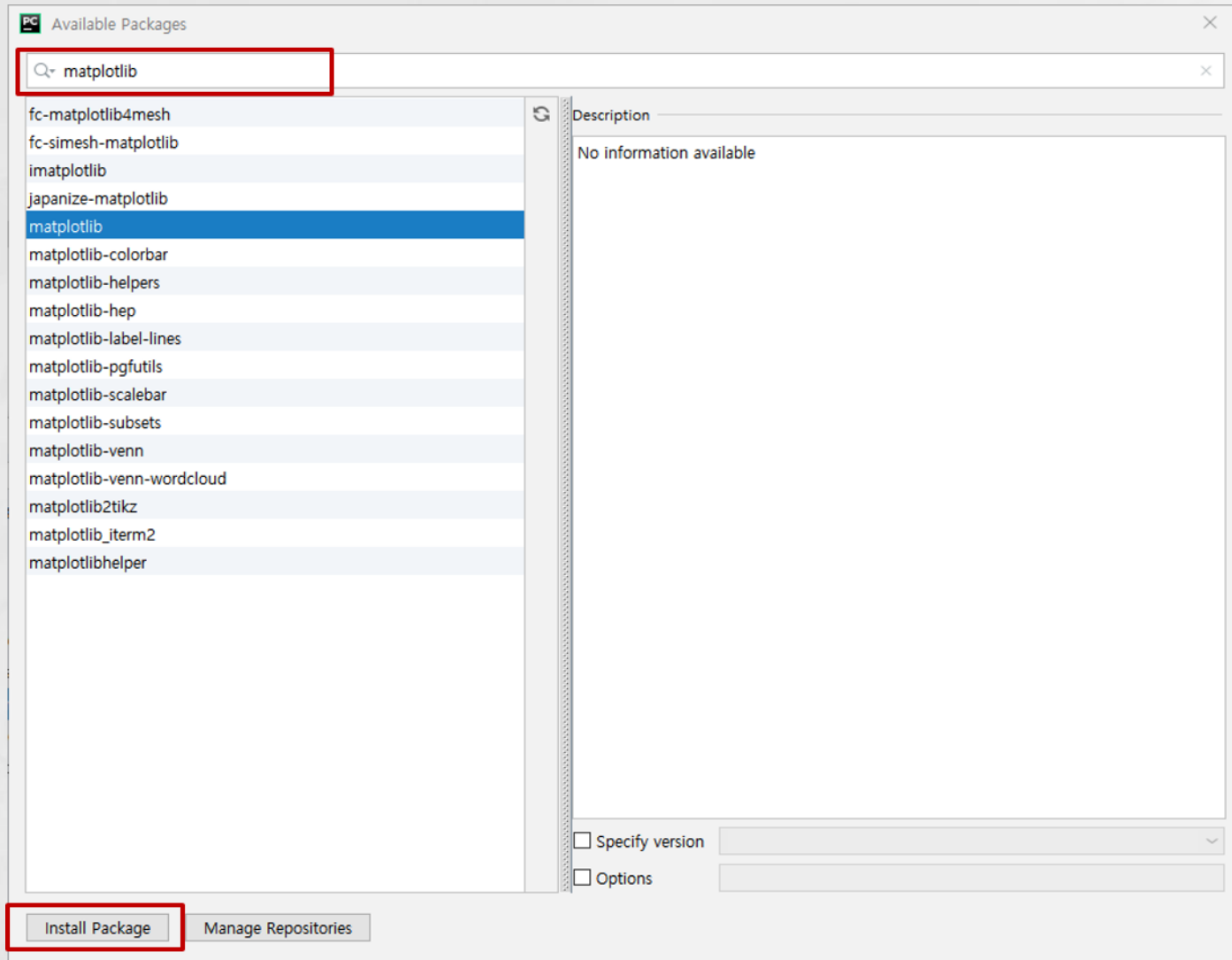


'div[class=thumb\_inner02] > dl[class=thumbnail\_table01]'

7개 칼럼으로 DataFrame 생성

```
<div class="thumb_inner02">
<dl class="thumbnail_table01">
<dt class="thumbnail_img02">
</dt>
<dd>서울-서초-2017-00092</dd>
<dt class="thumbnail_img02">
</dt>
<dd>2017-06-30</dd>
<dt class="thumbnail_img02"></dt>
<dd>기타축종</dd>
<dt class="thumbnail_img02"></dt>
<dd>미상</dd>
<dt class="thumbnail_img02"></dt>
<dd>반포동 두리동물병원..</dd>
<dt class="thumbnail_img02"></dt>
<dd>총19마리리빙박스예..</dd>
<dt class="thumbnail_img02"></dt>
<dd>종료(입양)</dd>
</dl>
</div>
```

# 수집 자료 시각화





## ● Top5 단어 시각화

