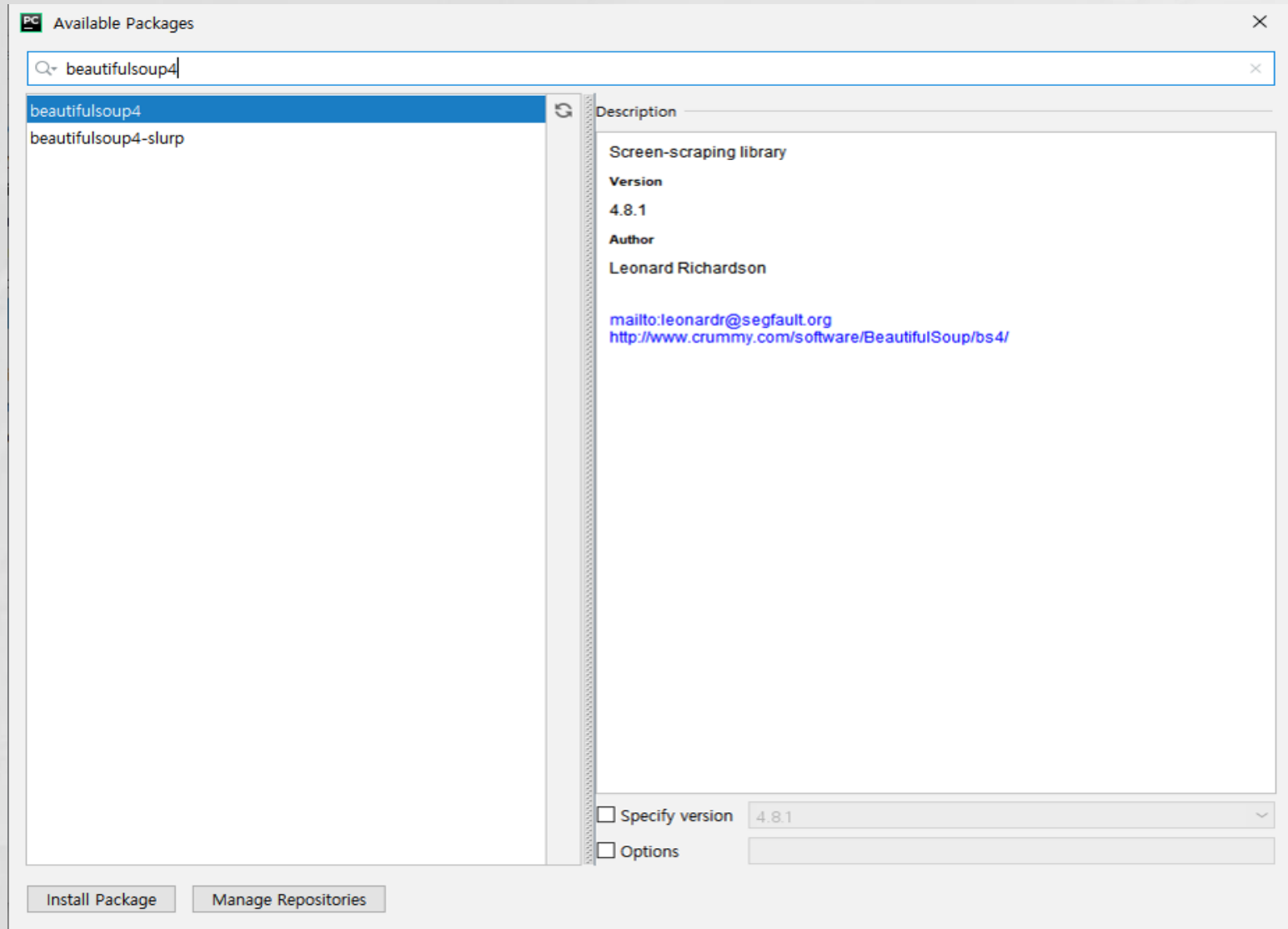


Chap08. Web Crawling

작성자 : 김진성

beautifulsoup4

Pycharm 패키지 설치



Settings



> Appearance & Behavior

Keymap

> Editor

Plugins

> Version Control

> Project: workspace

Python Interpreter

Project Structure

> Build, Execution, Deployment

> Languages & Frameworks

> Tools

Project: workspace > Python Interpreter

For current project

Python Interpreter: Python 3.7 C:\Python37\python.exe

Package	Version	
PyMySQL	0.0.2	
Werkzeug	1.0.1	1.0.1
beautifulsoup4	4.9.0	▲ 4.9.1
click	7.1.1	▲ 7.1.2
cycler	0.10.0	0.10.0
itsdangerous	1.1.0	1.1.0
kiwisolver	1.2.0	1.2.0
matplotlib	3.2.1	▲ 3.3.2
numpy	1.17.2	▲ 1.19.2
pandas	1.1.2	1.1.2
pip	20.1.1	▲ 20.2.3
pyparsing	2.4.7	2.4.7
python-dateutil	2.8.0	▲ 2.8.1
pytz	2019.1	▲ 2020.1
setuptools	47.1.0	▲ 50.3.0
six	1.12.0	▲ 1.15.0
soupsieve	2.0	▲ 2.0.1

설치된 패키지



OK

Cancel

Apply

Html Parsing Web Crawling

```
import urllib.request # url 요청 모듈  
from bs4 import BeautifulSoup # html 양식으로 파싱
```

```
# 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)  
src = data.decode("utf-8")  
html = BeautifulSoup(src, 'html.parser')
```

```
# 3) <a> 태그 수집  
a = html.find('a')
```



2. html의 <a>태그 가져오기

```
links = html.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
```

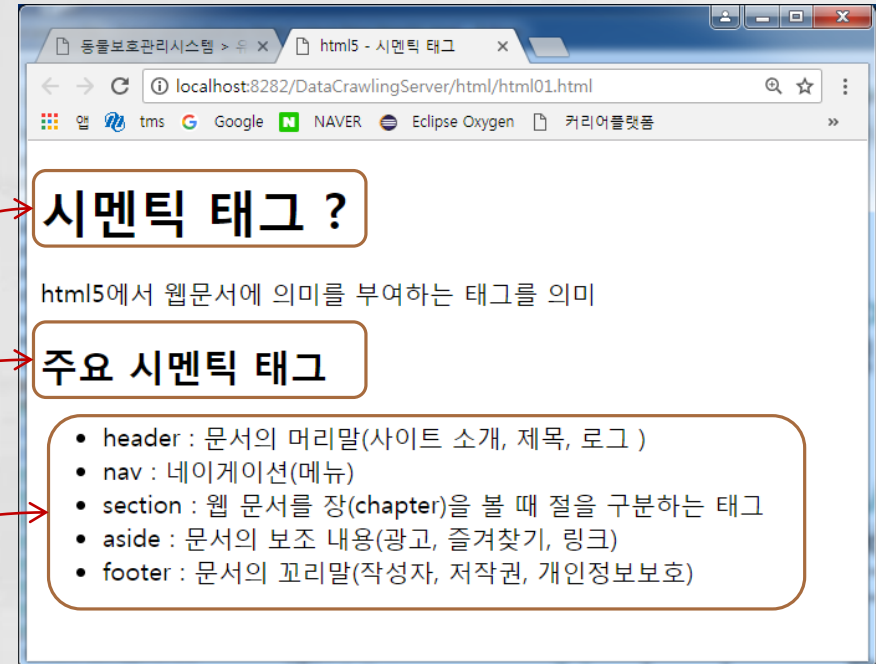
```
li = soup.find("li")
```

```
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 shows the homepage of the Animal Protection Center website. The header includes the site name '동물보호관리시스템' and navigation links like '정책홍보', '유기동물·동물보호센터', '동물등록', '농경동물', '상형동물', '자료매달', and '동물단체입/양요업'. The main banner features a girl and a dog with the text '소중한 생명! 여러분의 반려동물 사랑으로 지켜주세요.' Below the banner, there are links for '반려동물물려주고 싶으세요?', '입양안내책보기', '구조편경 온라인 신청', and '동물등록증 출력하기'. The middle section has three columns: '공지' (Notice) with a list of recent notices, '유기동물 공고' (Lost Animal Notice) with a photo of a dog, and '찾아주세요' (Find Your Pet) with a photo of a dog. The bottom section contains several informational tiles: '동물보호센터 검색' (Animal Protection Center Search), '반려동물을 잃어버리셨나요?' (Lost Your Pet?), '보호종 동물검색' (Protected Breed Animal Search), '동물등록제 전국 확대' (National Expansion of Animal Registration System), '찾아주는 동물' (Find Your Pet), and '동물대행입양 검색' (Animal Proxy Adoption Search).

[유기동물공고] 페이지

http://www.animal.go.kr/portal_rnl/abandonment/public_list.jsp

The screenshot displays the Animal Protection Management System (APMS) website. The main navigation bar includes links for 'Policy Information', 'Abandoned Animal Protection Center' (highlighted with a red box), 'Animal Registration', 'Domestic Animals', 'Wild Animals', 'Data Room', and 'Animal Sale/Adoption'. Below this, a sub-navigation bar lists 'Public', 'Lost and Found', 'Adoption', 'Animal Registration', 'Animal Protection', and 'Animal Welfare (TNR)'. The left sidebar contains a 'Public' menu with options like 'Lost and Found', 'Adoption', 'Animal Registration', 'Animal Protection', and 'Animal Welfare (TNR)'. The main content area features a banner for the 'Abandoned Animal Protection Center' and a section titled 'Public List' with a red box around the 'Public List' link. Below this, there is a search bar and a table of public lists. The table has columns for 'Date', 'Animal Name', 'Animal Type', 'Animal Status', and 'Animal Location'. The table contains one row with the following data: '2017-12-12', '2018-01-12', '2018-01-12', '2018-01-12', '2018-01-12'. The bottom of the page includes a footer with the website URL and a search bar.

동물보호관리시스템
ANIMAL PROTECTION MANAGEMENT SYSTEM

로그인 | 회원가입 | 아이디/패스워드찾기 | 사이트맵 | 등록번호 | 등록번호 15자리

정책정보 | 유기동물·동물보호센터 | 동물등록 | 농장동물 | 실험동물 | 자료마당 | 동물판매업/장묘업

공고 | 분실신고 | 습득시 안내 | 입양안내 | 유기동물보호센터 | 보호중동물 | 길고양이중성화(TNR)

— 공고
— 분실신고
— 습득시 안내
— 입양안내
— 유기동물보호센터
— 보호중동물
— 길고양이 중성화(TNR)

동물등록번호 15자리를 입력하여
동물의 이름, 성별, 품종, 관할기관 등을
검색할 수 있으며,
관할기관에 문의하여 유기동물의 소유자를
찾을 수 있습니다.

선택

검색

유기동물·동물보호센터
Animal Protection Management System

유기동물 공고

HOME > 유기동물 동물보호센터 > 공고 > 유기동물공고

유기동물 개요 | 유기동물 공고

유기동물공고

「동물보호법」 제 17조 및 동법 시행규칙 제 7조에 따라 유기·유실동물을 보호하고 있는 경우에는 소유자 등이 보호조치 사실을 알 수 있도록 7일 동안 공고하여야 합니다.
공고중인 동물 소유자는 해당 시·군·구 및 동물보호센터에 문의하여 동물을 찾아가시기 바랍니다.
다만, 「동물보호법」 제19조 및 동법 시행규칙 제21조에 따라 소유자에게 보호비용이 청구될 수 있습니다. 또한 「동물보호법」 제17조에 따른 공고가 있는 날부터 10일이 경과하여도 소유자 등을 알 수 없는 경우에는 「유실물법」 제12조 및 「민법」 제253조의 규정에도 불구하고 해당 시·도지사 또는 시장·군수·구청장이 그 동물의 소유권을 취득하게 됩니다.

시·도지사, 시장·군수·구청장(직인 생략)

SEARCH 날짜 입력시 다음 예와같이 입력해 주세요 예) 2011-01-01
날짜 2017-12-12 ~ 2018-01-12 (날짜는 필수일 기준입니다)

www.animal.go.kr/portal_rnl/abandonment/public_list.jsp

동물등록제 전국 확대 시행
동물등록제 전국 확대 시행
동물등록제 전국 확대 시행

동물복지 축산농장 인증제 안내
동물보호복지 온라인 교육

동물등록 모바일 서비스 이용안내

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 콤마를 기준으로 합니다)
 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 입양희망 동물</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, 검게심어 ..</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 동해..</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>공고번호 전남-순천-2018-00023</p>	 <p>공고번호 전남-순천-2018-00022</p>		

동물등록제도 전국 확대 시행

동물복지 확산을 위한 정책 안내

동물보호법제 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 : 품종유기가 발생할 수 있으나 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 닥스훈트
 성별 암컷
 발견장소 강남구청
 특징 장모종. 코갈색.유선..
 상태 공고중



자세히 보기

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



자세히 보기

공고번호 서울-강남-2018-00003
 접수일 2018-01-03
 품종 푸들
 성별 수컷
 발견장소 도곡동 416-7..
 특징 백내장.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 노물자국.배꼽할증..
 상태 공고중



자세히 보기

공고번호 서울-강남-2017-00243
 접수일 2017-12-18
 품종 닥스훈트
 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 빨간바탕에노란팔2개.. 상태 종료(반환) 자세히 보기	

전체 검색 페이지 48 페이지[현재 : 48page]

40 41 42 43 44 45 46 47 48 49

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 생략

전국 확대시행

SEARCH

날짜 입력시 다음 예와같이 입력해주세요 예) 2011-01-01
 날짜 2015-01-01 ~ 2018-01-12 (날짜는 필수입력 기준입니다)
 시도 서울특별시 선택 시군구 선택 보호센터 전체
 종류 전체 선택 상태 전체 조회

※ 검색시 유의사항 : 풍중오류가 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 치석있고 부절 교합이며...
 상태 중요(반한)

자세히 보기

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

자세히 보기

공고번호 서울-종산-2015-00002
 접수일 2015-01-01
 품종 말티즈
 성별 수컷
 발견장소 이촌아파트 중간 도...
 특징 눈물흘리고있음
 상태 중요(반한)

자세히 보기

2621 2622 2623 2624 2625 2626 2627 2628

이용안내 | 개인정보처리방침 | 저작권 정책
 (우)39660 경상북도 김천시 학신8로 177(출곡동) 업무분장: 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

동물등록 모바일 서비스
 이용안내

WA
 WEB ACCESSIBILITY
 KOREAN GOVERNMENT

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
copyright by Animal and Plant Quarantine Agency. All Rights Reserved.



유기견 자료 Crawling 대상 문서

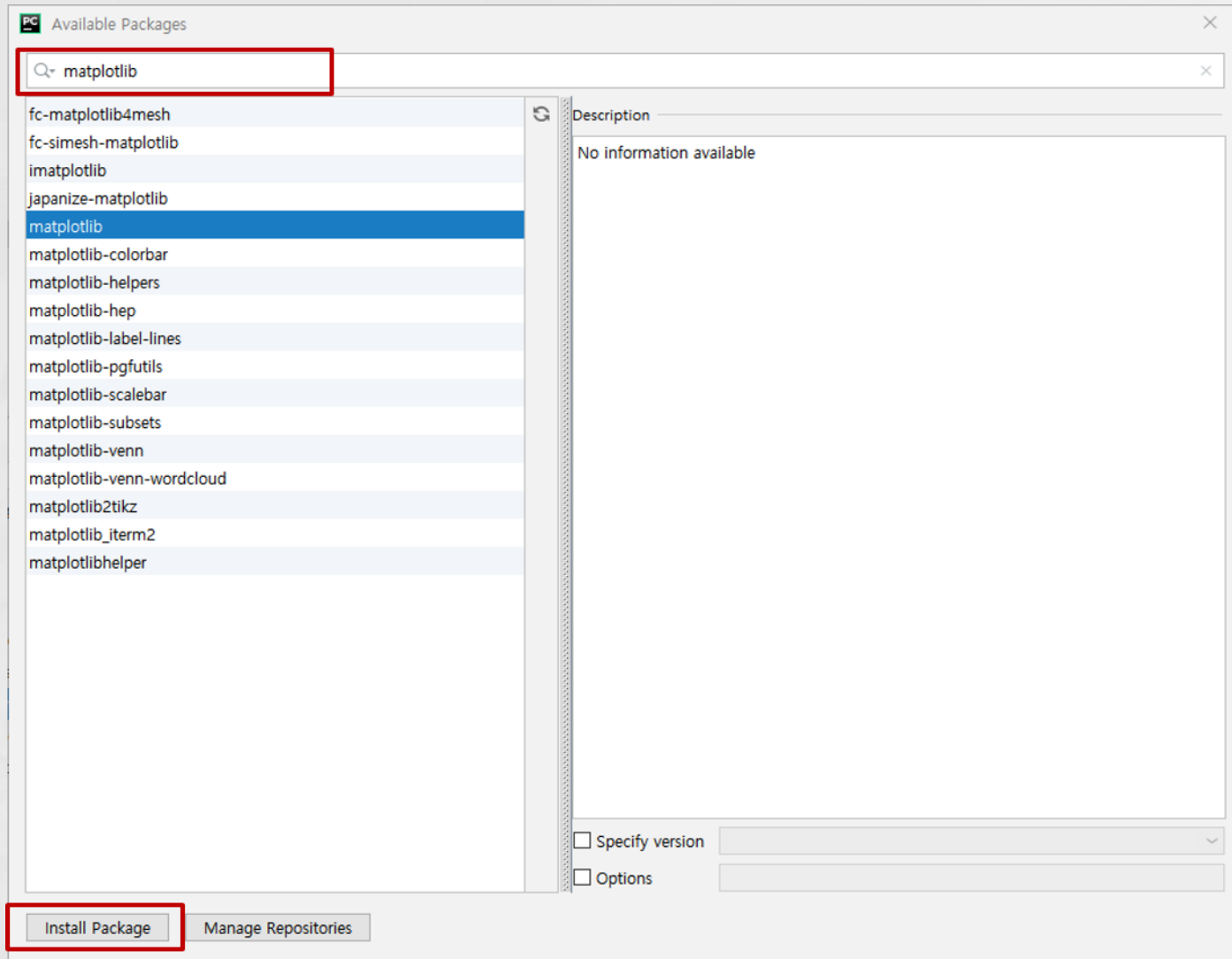


'div[class=thumb_inner02] > dl[class=thumbnail_table01]'

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

