
AI파이썬 빅데이터수집

정적웹페이지 크롤링

크롤링:크롤링(crawling) 혹은 스크레이핑(scraping)은 웹 페이지를
그대로 가져와서 거기서 데이터를 추출해 내는 행위다

크롤링
- 텍스트(HTML)

빈도분석

최종목표

정적웹페이지 크롤링

- 정적 웹페이지 크롤링

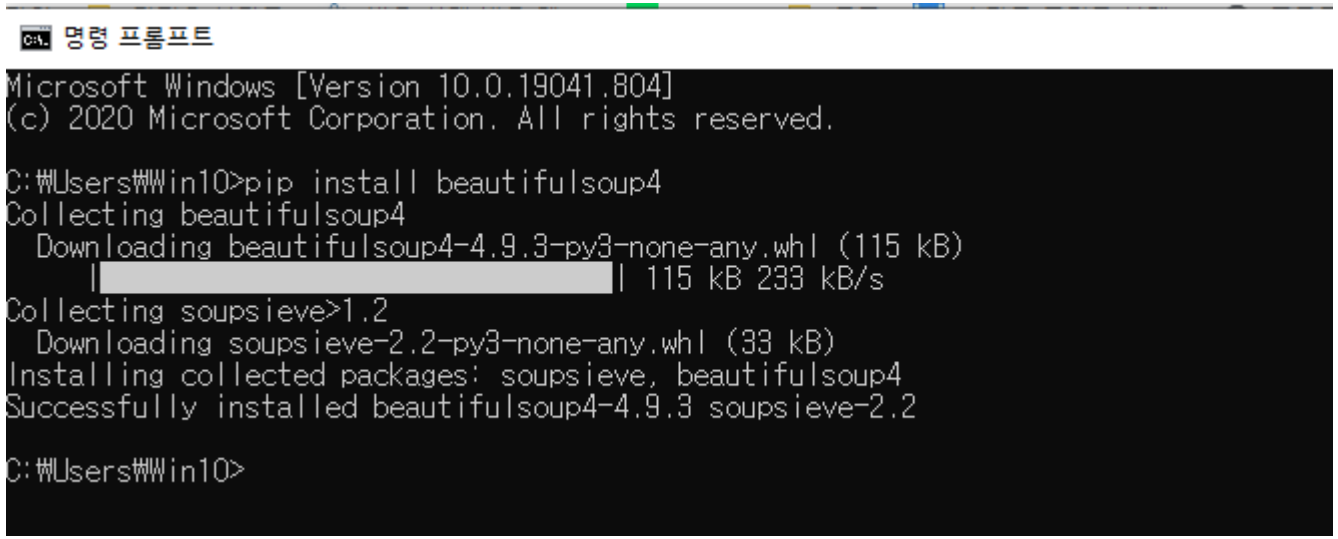
1. HTML 구조를 분석한 뒤 필요한 데이터를 직접 크롤링 해야한다.
2. 웹페이지의 HTML 구조를 분석하는 작업을 파싱(parsing) 이라한다.
3. 파이썬 을 이용한 라이브러리 BeautifulSoup 사용.

정적웹페이지 크롤링

- 파이썬 작업

1. BeautifulSoup 설치

pip install beautifulsoup4



정적웹페이지 크롤링

- 파이썬 코딩

  beautifulSoup 사용법

```
import requests
from bs4 import BeautifulSoup as bs

tag = "<p class='big_data01' id='uniq01'> Hello World! </p>"
soup = bs(tag,'html.parser')

# 태그 이름만 특정
data01 = soup.find('p')
print(data01)

data02 = soup.find('p').text
print(data02)

# 태그 속성만 특정
data03 = soup.find(class_='big_data01')
print(data03)
data04 = soup.find(attrs = {'class':'big_data01'})
print(data04)

# 태그 이름과 속성 모두 특정 (class로 검색)
data05 = soup.find('p',{'class':'big_data01'}).text
print(data05)
data06 = soup.find('p',{'class':'big_data01'}).text
print(data06)

# 태그 이름과 속성 모두 특정 (ID로 검색)
data07 = soup.find('p',{'id':'uniq01'}).text
print(data07)
```

정적웹페이지 크롤링

- 파이썬 코딩

네이버 미세먼지 크롤링

```
import requests
from bs4 import BeautifulSoup as bs

url = 'https://search.naver.com/search.naver?query=날씨'
html = requests.get(url)
soup = bs(html.text,'html.parser')

data1 = soup.find('div',{'class':'detail_box'})
#print(data1)

data2 = data1.findAll('dd')
#print(data2[0])

#미세먼지
dust01 = data2[0].find('span',{'class':'num'}).text
print(dust01)

#초미세먼지
dust02 = data2[1].find('span',{'class':'num'}).text
print(dust02)

#오존
o3 = data2[2].find('span',{'class':'num'}).text
print(o3)

#온도
data3 = soup.find('div',{'class':'main_info'})
#print(data3)
temp01 = data3.find('span',{'class':'todaytemp'}).text
print(temp01)
```

정적웹페이지 크롤링

- 파이썬 코딩

다음 날씨 크롤링

```
import requests
from bs4 import BeautifulSoup as bs


html = requests.get('https://search.daum.net/search?nil_suggest=btn&w=tot&DA=SBC&q=%EC%9D%B8%EC%B2%9C%EA%B4%91%EC%97%AD%EC%8B%9C+%EB%8F%99%EA%B5%AC+%EB%82%A0%EC%94%A8')
soup = bs(html.text,'html.parser')

data1 = soup.find('div',{'class':'info_weather'})
print(data1)

temp = data1.find('strong',{'class':'txt_temp'}).text
print(temp)
```