

빅데이터 수집 시스템 개발

김용선

INDEX

01

수집 데이터 분석

02

수집에 사용한 코드

03

수집 데이터 확인

01. 수집 데이터 분석

서울


곤지암리조트점 ①
경기 광주시 도척면 도용리 96-11 ②
031-798-0987

조지점
경기 안산시 단원구 조지로 114 (조지동, 주공프라자)
031 403 5988

죽전점
경기 용인시 수지구 죽전동 1175-9
031-889-7030

동탄방교점
경기 화성시 동탄기흥로247번길 10-23 (방교동)
031-376-7999

수서점
서울 강남구 광평로47길 17 (수서동, 신동아아파트)



```
<div id="section">
  <!-- 내용 -->
  <div class="fstoreArea">
    <div class="tabArea">...</div>
    <div class="fssearch wo">...</div>
    <div class="mapArea">
      <div class="maptxt">
        <ul>
          <li>
            <div class="cityarea">
              <div class="citytit">서울</div>
              <div class="locarea active" style>
                <ul>
                  <li class="storeLI block active" data-name="곤지암리조트점" == $0
                    <a href="javascript:void(0);" data-idx="2101">
                      <p class="loc_tit">곤지암리조트점 ①
                      <p class="loc_lo" data-coordx="37.3328194752" data-coordy="127.3107298275">경기 광주시 도척면 도용리 96-11</p> ②
                      <p class="loc_call">031-798-0987</p>
                    </a>
                    <div class="storeImg" style="display: block;">...</div>
                  </li>
                  <li class="storeLI block" data-name="조지점">...</li>
                  <li class="storeLI block" data-name="죽전점">...</li>
                  <li class="storeLI block" data-name="동탄방교점">...</li>
                  <li class="storeLI block" data-name="수서점">...</li>
                  <li class="storeLI block" data-name="역삼점">...</li>
                  <li class="storeLI block" data-name="압구정점">...</li>
                </ul>
              </div>
            </li>
          </ul>
        </div>
      </div>
    </div>
  </div>
</div>
```

수집하려는 데이터 구조 확인

02. 수집에 사용한 코드_01

```
from selenium import webdriver
from bs4 import BeautifulSoup
import time
import storedb

# 웹드라이버 객체 생성
driver = webdriver.Chrome('./webdriver/chromedriver')
driver.implicitly_wait(2)

# url에 접근
driver.get('https://www.mexicana.co.kr:50010/company/find_store.asp')
time.sleep(3)

# driver에서 페이지 소스를 가져옴
html = driver.page_source
# html 부분을 파싱
soup = BeautifulSoup(html, 'html.parser')
```

02. 수집에 사용한 코드_02

```
# 필요한 데이터들을 감싸고 있는 태그를 가져옴
storelist = soup.select('div.locarea ul .storeLI > a')

storeNameList = [] # db에 저장할 지점들의 이름을 담은 리스트
storeAddrList = [] # db에 저장할 지점들의 주소를 담은 리스트

# 파싱해온 태그의 개수만큼 반복하여 필요한 데이터 추출
for store in storelist:
    name = store.select('.loc_tit')[0].text
    addr = store.select('.loc_lo')[0].text
    storeNameList.append(name)
    storeAddrList.append(addr)

# 이름과 주소를 담은 리스트들을 db에 저장
storedb.insert_data(storeNameList, storeAddrList)
```

02. 수집에 사용한 코드_03

```
def insert_data(name, addr):  
    # name : db에 담을 지점 이름들이 담긴 리스트  
    # addr : db에 담을 지점 주소들이 담긴 리스트  
    conn = oci.connect(connContent)  
    cursor = conn.cursor()  
    sql = """  
        INSERT INTO STORE(NAME, ADDR)  
        VALUES(:name, :addr)  
    """  
    for i in range(len(name)): # name 리스트의 길이만큼 반복하여 db에 저장  
        cursor.execute(sql, (name[i], addr[i]))  
    cursor.close()  
    conn.commit()  
    conn.close()
```

03. 수집한 데이터 확인

```
select * from store;
```

질의 결과 x

SQL | 인출된 모든 행: 860(0.073초)

NAME	ADDR
843 표선점	제주특별자치도 서귀포시 표선면 표선동서로 253
844 김녕점	제주특별자치도 제주시 구좌읍 김녕로 122
845 터미널점	제주특별자치도 제주시 남성로 41 (용담1동)
846 일도점	제주특별자치도 제주시 동광로20길 29 (일도2동)
847 화북점	제주특별자치도 제주시 동화로 3 (화북1동)
848 멕시코나도련점	제주특별자치도 제주시 매촌1길 18 (도련2동)
849 제대점	제주특별자치도 제주시 산천단동3길 29 (아라1동)
850 용담점	제주특별자치도 제주시 서사로 10 (용담1동)
851 하귀해안로점	제주특별자치도 제주시 애월읍 가문동남길 63 (하귀2리)
852 하귀해안로점	제주특별자치도 제주시 애월읍 가문동남길 63 (하귀2리)
853 고성하귀점	제주특별자치도 제주시 애월읍 고성서3길 15-5 (제주고성휴먼시아)
854 애월곽지점	제주특별자치도 제주시 애월읍 일주서로 6001
855 법원점	제주특별자치도 제주시 오복2길 31-14 (이도2동)
856 외도점	제주특별자치도 제주시 외도일동 486-12
857 노형점	제주특별자치도 제주시 월랑북길 31 (노형동)
858 연동점	제주특별자치도 제주시 은남4길 37 (연동)
859 한라대점	제주특별자치도 제주시 정원로 53-1 (노형동)
860 첨단점	제주특별자치도 제주시 첨단로동길 106 (월평동)

THANK YOU!