SQL, 웹 크롤링을 활용한 전국 직업 훈련 교육 현황 파악

경기대학교
한광종

Contents

0 1 / 주제 선정 및 목적

① 서울, 수도권 지역에 밀집되어 있는 직업 훈련 교육 현황을 파악

② 비 수도권 지역의 교육 훈련이 부족하다는 근거를 구체적으로 제시

③ 3주차에 배운 DB Query를 실습

0 2 / 데이터 수집 - 1

직업훈련포털
→ 웹 크롤링(Python)

시도, 성별 실업자수
(KOSIS 17년4분기~)

시도군 연령별 인구 (KOSIS,201806)

지도 관련 좌표 데이터

0 2 / 데이터 수집 - 3

① 크롤링할 내용 정함(범위 포함)

② Firefox App ‘Firebug’로 path와 숨겨진 URL 얻음

③ Python Selenium, webdriver 모듈로 얻어온 path와 반복문을 사용하여 텍스트를 얻어 Oracle DB(Localhost)에 저장

※ 약 50,300개의 직업 훈련 과정에 대한 데이터(과정 명, 훈련기간, 훈련시간 등)를 정제하지 않고 그대로 저장

0 3 / 데이터 탐색 및 전처리 - 1

① Python 모듈 cx_Oracle을 이용하여 Jupyter Notebook 환경에서 Query 작업, 데이터 정제

② CSV 파일(KOSIS 데이터)를 Oracle DB에 업로드(정제된 자료로 용량이 적어 EXCEL로 간단한 작업)

③ 정제된 크롤링 데이터와 CSV 파일을 시군구 명을 기준으로 합침

ADDRESSTEL
(경기 수원시 팔달구 ☎ 031-243-2111)
(서울 구로구 ☎ 02-2631-3139)
(경북 구미시 ☎ 054-462-3610)
(서울 동대문구 ☎ 02-957-0915)
(대구 달서구 ☎ 053-581-1231)
(대구 중구 ☎ 053-255-5588)
(경기 부천시 ☎ 032-325-7707)
(대구 북구 ☎ 053-954-8286)
(서울 성동구 ☎ 02-461-1101)

```
In [111]: #ADDRESSTEL -> FULL_ADDRESS
full_address_cleansing= """SELECT NVL(RTRIM(SUBSTR(REPLACE(REPLACE(ADDRESSTEL, ' )')
                          ,'( ',''),1,REGEXP_INSTR(REPLACE(REPLACE(ADDRESSTEL, ' )'),'( ',''),' ☎ ')),NULL)
                          AS FULLADDRESS
                          """
```

```
import cx_Oracle
```

```
con =  cx_Oracle.connect('hr/password@localhost:1521/xe')
cur = con.cursor()
```

```
create_table = """CREATE TABLE HRD_NET(
                    NO NUMBER(38),
                    COURSE_NAME VARCHAR2(500),
                    ACADEMY VARCHAR2(500),
                    FULL_ADDRESS VARCHAR2(500),
                    SIDO VARCHAR2(500),
                    SIGUNGU VARCHAR2(500),
                    TEL VARCHAR2(500),
                    START_DATE VARCHAR2(500),
                    END_DATE VARCHAR2(500),
                    EMPLOY_SCORE NUMBER(30),
                    EMPLOY_WAGE NUMBER(30),
                    TRAIN_PRICE NUMBER(30),
                    RECRUIT_LIMIT NUMBER(30),
                    WEEKEN_ABLE VARCHAR2(500),
                    TRAIN_PERIOD_DAY NUMBER(30),
                    TRAIN_PERIOD_TIME NUMBER(30),
                    COUR_SATISFACTION NUMBER(30),
                    OCCUPATION VARCHAR2(500),
                    TRAIN_LEVEL NUMBER(10),
                    NCS VARCHAR2(500),
                    RELEATED_CERTIFICATE VARCHAR2(500)
                    )"""
cur.execute(create_table)
```

0 4 / 데이터 탐색 및 전처리 - 2

크롤링한 데이터

NO	COURSE_N	ACADEMY	FULL_ADDRESS	SIDO	SIGUNGU	TEL	START_DAT	END_DATE	EMPLOY_S	EMPLOY_V	TRAIN_PRI	RECRUIT_I	WEEKEN_A	TRAIN_PEI	TRAIN_PEI	COUR_SAT	OCCUPATI	TRAIN_LE	NC
42190	사무행정_	비아이티컨	강원 춘천시	강원	춘천시	033-263-4	2018-04-18	2018-06-19	40	1930731	669480	15	wkdays	40	120	100	사무행정	3	YE
42164	굴삭기기능	창원중장비	경남 창원시 마산회원구	경남	창원시	055-292-0	2018-04-18	2018-07-04	58	2346354	1508420	20	wkdays	52	208	80	굴삭기운전	2	YE

시도별 좌표 정보 데이터

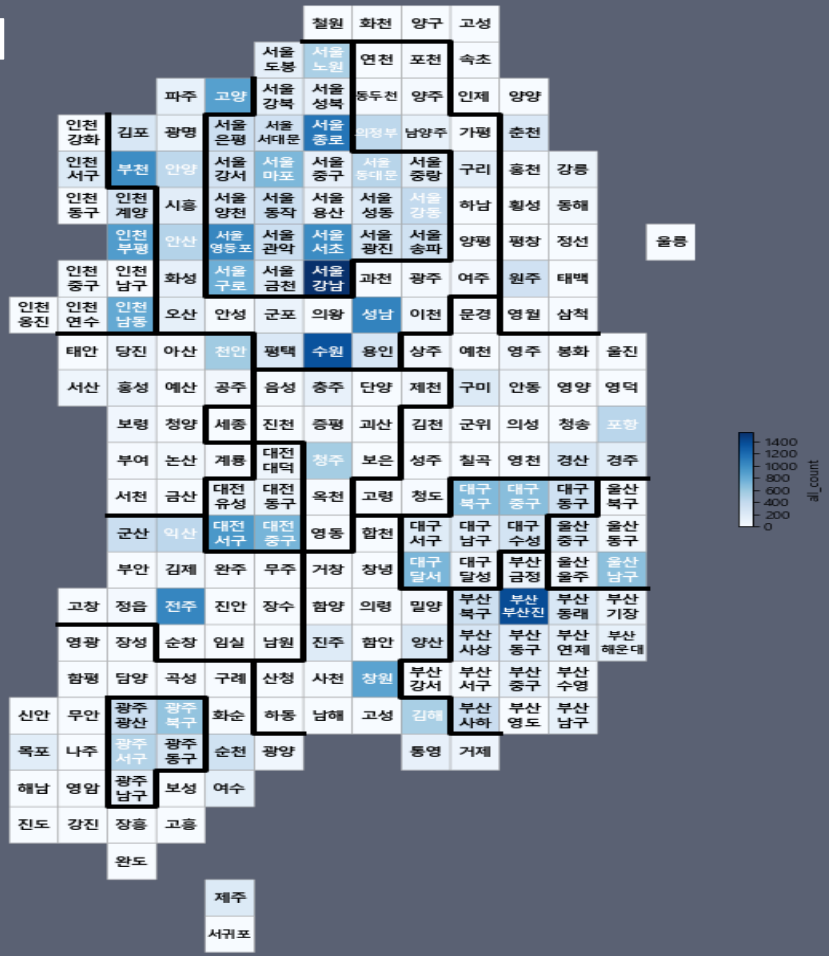
	A	B	C	D	E	F	G	H
	index	inguNumb	shortNam	x	y	ground	sido	sigu
	0	202520	강릉	11	4	1040.07	강원도	강릉시
	1	25589	고성(강원)	9	0	664.19	강원도	고성군
	2	86747	동해	11	5	180.01	강원도	동해시
	3	63986	삼척	11	8	1185.8	강원도	삼척시

시군구별 인구 데이터

	A	B	C
1	sido	sigungu	ingu
2	전체	전체	51790131
3	서울특별시	전체	9830452
4	서울특별시	종로구	154312
5	서울특별시	중구	125986
6	서울특별시	용산구	229939
7	서울특별시	성동구	307460
8	서울특별시	광진구	356757

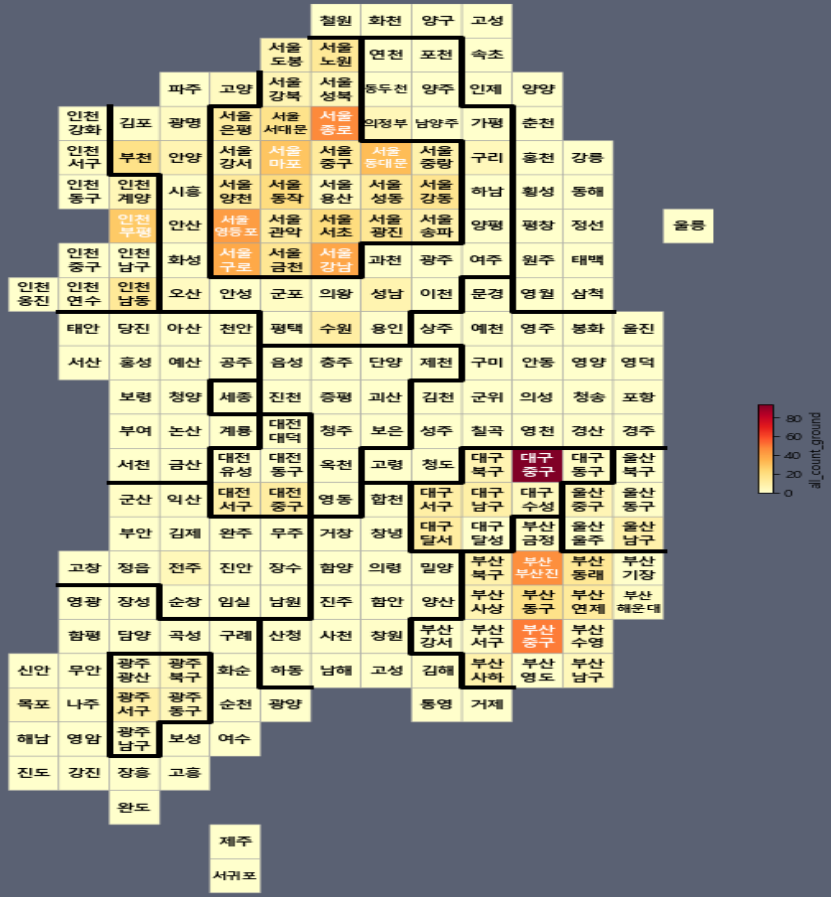
분석 데이터

0 4 / 데이터 시각화 -1



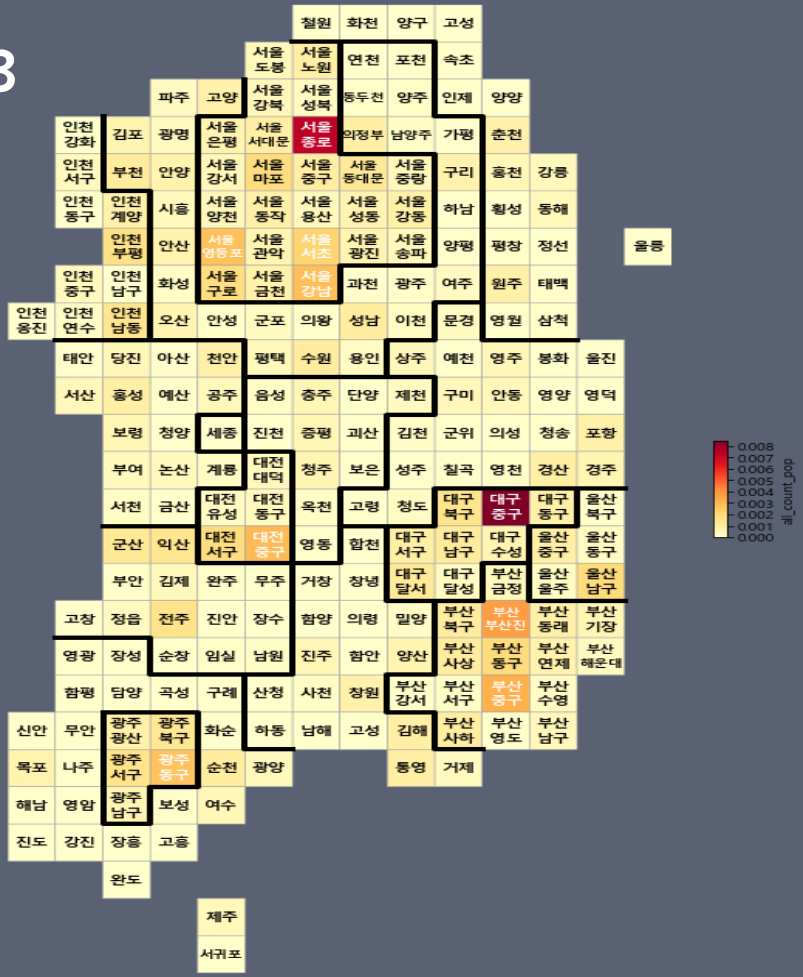
지도1) 지역 별 전 분야 교육 수

0 4 / 데이터 시각화 -2



지도2) 지역 별 면적 대비 전체 교육 수

0 4 / 데이터 시각화 -3



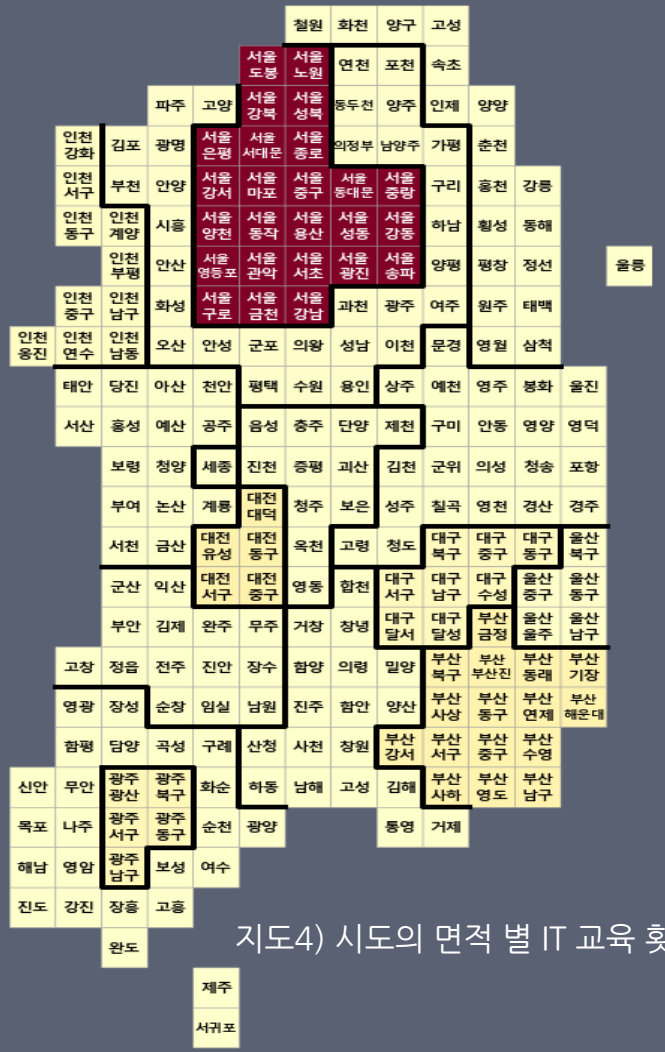
지도3) 인구 수 별 전체 교육 횟수

0 4 / 결론

※ IT 교육의 분류는 직업 훈련 포털 기준, 업종 분류를 기준으로 함

'응용SW엔지니어링','빅데이터',
'SW제품기획', '빅데이터플랫폼구축', 'IT기술지원',
'IOT융합서비스기획', 'SW아키텍처', '정보기술전략'

※ 시도의 면적 별 IT 교육 횟수를 비교 했을 때, 서울 지역의 교육 횟수 비중이 압도적으로 높은 것을 알 수 있음.



지도4) 시도의 면적 별 IT 교육 횟수

감사합니다