|  |  |
| --- | --- |
| 교육 제목 | **Pandas - 데이터프레임과 웹 스크래핑** |
| 교육 일시 | 21년 09월 29일 |
| 교육 장소 | YGL - 비대면 수업 (자택) |
| **교육 내용** | |
| 오전 | 데이터프레임  열 선택: 객체['인덱스명'], 객체[['인덱스1', '인덱스2...']] 인덱스 이름이면 따옴포''/”” 를 붙이고 인덱스 위치면 그냥 숫자를 쓴다.  행과열선택: 객체.loc/iloc[[행인덱스],[열인덱스]] / 객체.loc/iloc[[행인덱스1,행인덱스2],[열인덱스1,열인덱스2]] 범위를 지정하면 객체.loc/iloc['인덱스명':'인덱스명2']나 [0:-1] 로 대괄호를 하나만 쓰지만, 하나하나 지정하려면 객체.loc/iloc[['인덱스','인덱스3','인덱스5']] 이런식으로 대괄호를 2개를 입혀 리스트로 만들어야 한다.  -열 추가: 객체['추가 열 이름'] = 데이터 값 (열은 무조건 값만 들어올 수 있다)  -행 추가: 객체.loc['추가 행 이름'] = 데이터 값/배열 ex) df.loc[”new row] = ['new element', 100, 500, 900] -> 배열로 new row 행에 값을 입력함.  -행,열 위치 바꾸기 : df.transpose()나 df.t 를하면 열과 행이 서로 바뀐다.  -행 인덱스 재배열: 객체.reindex(새로운 인덱스 배열)로 행 인덱스를 추가 할 수 있다, 하지만 초기값이 Nan으로 채워지기 때문에 주의 필요하고 ,fill\_value=0으로하면 Nan대신 0으로 값이 대체된다.  -행 인덱스 초기화: 객체.reset\_index()로 정수형 위치 인덱스로 초기화 하고 기존 행 인덱스는 열로 이동하고 새로운 데이터프레임 객체를 반환한다.  -행 인덱스를 기준으로 데이터프레임 정렬: 객체.sort\_index(ascendig=True/False)로 index를 ascending 오름차 순으로 정렬해주는데 False로 값을 넣어주면 내림차순으로 정령된다. 그리고  -객체.sort\_value ('특정 열', ascendig=True/False)로 특정 열의 데이터 값을 기준으로 데이터 프레임을 정렬할 수 있다.  -판다스 객체의 산술연산:3단계의 처리 과정을 거치는데 1. 행/열 인덱스를 기준으로 모든 원소를 정렬 2. 동일한 위치에 있는 원소끼리 일대일로 대응 3.일대일 대응이 되는 원소끼리 연산을 처리한다. 이때 대응 되는 원소가 없으면 NaN 값으로 대체된다. |
| 오후 | 데이터 입출력  -외부파일 읽기  \*csv - comma-separated values 라고 부르틑 텍스트 파일이고 쉼표로 열을 구분하고 줄바꿈으로 행을 구분한다. 읽기는 read\_csv이고 파일은 2차원 배열로 데이터프레임으로 변경해서 사용한다. 보내기/저장은 객체.to\_csv(파일명)  \*json -JavaScript에서 유래한 데이터 공유를 목적으로 개발된 특수한 파일형식이고 파이썬의 딕셔너리와 비슷하고 key : value 구조를 갖고있고, read\_json/ to\_json 로 읽고 저장한다.  \*html - read\_json/to\_json  \*local clipboard - read\_clipboard/ to\_clipboard  \*ms excel -우리가 흔히 알고 있는 Microsoft에서 개발한것으로 엑셀의 행과 열은 데이터프레임의 행과열에 일대일 대응된다. csv랑 같이 read\_excel 로 읽고 to\_excel로 보낸다  \*hdf5 format -read\_hdf / to\_hdf  \*sql - read\_sql / to\_sql  웹 스크래핑: BeautifulSoup 과 같은 웹 스크래핑 도구로 웹에서 데이터를 가져오는 것.  API를 이용한 데이터 수집 :구글지오 같은 api나 네이버 파파고 api같은 것들이 있는데 key를 요구하면 거기서 말하는데로 key를 획득해서 입력해야 사용가능하다.  오늘 강의 간단히 정리  df = pandas.DataFrame ....  행\_df.loc[인덱스명], df.iloc[인덱스첨자], 여러개면 df.iloc[시작:끝]  열\_df.열 이름 , df[열이름]  원소 df.loc[[인덱스명]], [컬럼명]]  df.loc[[인덱스첨자],[컬럼의 순서-정수]] |