|  |  |
| --- | --- |
| 교육 제목 | **Seaborn, Folium 라이브러리** |
| 교육 일시 | 2021년 11월 10일 |
| 교육 장소 | 영우글로벌 강의실 |
| **교육 내용** | |
| 오전 | 1. Seaborn은 matplotlib의 기능과 스타일을 확장한 파이썬 시각화 도구의 고급 버전이다.   파이썬 실행 파일에 Seaborn 라이브러리를 임포트할때는 ‘sns’라는 약칭을 주로 사용한다.  -교재p147~p161   1. Folium 라이브러리   Folium 라이브러리는 지도 위에 시각화할 때 유용하다. 라이브러리 설치를 위해서 아나콘다 프롬프트로 설치해준다.  -교재p161~p169   1. 누락 데이터 처리 : 데이터를 파일로 입력할 때 빠트리거나 파일 형식을 변환하는 과정에서 데이터가 소실되는 것이 주요 원인이다. 일반적으로 유효한 데이터 값이 존재하지 않는 누락 데이터를 NaN(Not a Number)으로 표시한다.   -누락 데이터의 개수를 확인하려면 반드시 dropna=False 옵션을 사용해준다. 그렇지 않으면 NAN값을 제외하고 유효한 데이터의 개수만을 구한다.   1. 중복 데이터 처리 : 동일한 대상이 중복으로 존재하는 것은 분석 결과를 왜곡하기 때문에 중복 데이터를 찾아서 삭제해야한다. 중복 데이터를 제거하는 명령에는 drop\_duplicates()메소드가 있고, 원본 객체를 변경하려면 inplace=True 옵션을 추가한다. 2. 데이터 표준화 : 여러곳에서 수집한 자료들은 단위 선책, 대소문자 구분, 약칭 활용 등 여러가지 원인에 의해 다양한 형태로 표현된다. 이처럼 동일한 대상을 표현하는 방법에 차이가 있으면, 분석의 정확도는 현저히 낮아진다. 따라서 데이터 포맷을 일관성 있게 표준화 해야한다. |
| 오후 | 1. 범주형 데이터 처리   구간분할-연속변수를 일정한 구간으로 나누고, 각 구간을 범주형 이산 변수로 변환하는 과정  더미변수-카테고리를 나타내는 범주형 데이터를 회귀분석 등 머신러닝 알고리즘에 바로 사용할 수 없는 경우가 있는데, 컴퓨터가 인식 가능한 입력값으로 변환해야한다. 숫자 0, 1로 표현되는 더미 변수 이용   1. 정규화 2. 시계열 데이터   주식, 환율 등 금융 데이터를 다루기 위해 시계열 데이터를 쓴다.  특정한 시점을 기록하는 Timestamp와 두시점 사이의 일정한 기간을 나타내는 Period가 있다. |