데이터 시각화 프로젝트 학습 기록 정리

프로젝트 진행상황 기록 일지

6/14, 6/15: 깃허브 가입 및 설치와 사용방법 숙지, 도커설치, 프로젝트 해석, 파악 (결국 그래프 6개를 파이썬을 통해 구현하는 과정이 핵심), 데이터 세트 다운, 코드 분석 및 본문해석

6/16: MATPLOTLIB 기능 학습, 구글 COLAB 이용, 데이터분석용 파이썬 기초 코드 실습 및 정리, 데이터 전처리에 관해 학습 및 실습, COLAB과 깃허브를 연동시킴과 더불어 COLAB과 KAGGLE을 코딩을 통해 서로 연동하여 프로젝트에 필요한 데이터셋을 구글 드라이브에 저장하여 이용가능하게 하였음. (이로 인해 PC 저장공간을 굳이 쓸 필요 없게 만들어서 저장공간 효율성 증대) 이외에도 기본적인 코드 연습, 강의 수강 등 하루일과 중 대부분의 시간을 프로젝트 준비에 할애하였음.

6/17: PANDAS, NUMPY 관련 기본 학습 및 실습, 깃허브 목차 정리, 행렬 개념 심화학습, 사전과제 프로젝트 데이터 전처리 및 시각화 과정 진행

6/18: 최종 마무리 및 학습 기록 정리 및 검토, 시각화 코드 주석 보완

학습한 내용과 과정 요약 기록

파이썬 특징 정리

1. 언어들 중 배우기가 쉬운 편에 속하고.
2. 데이터를 다루는데 용이하며.
3. 다양한 모듈/패키지가 존재하여 활용범위가 넓다는 것!

(데이터 전처리의 주요 기법- 정제,통합,축소,변환)

데이터 분석을 하는데 파이썬이 필요한 이유

1. 파이썬은 데이터분석, 머신러닝/인공지능을 지원하는 다양한 모듈을 보유하고 있다.
    - pandas : 데이터분석을 쉽게 해주는 엑셀 파이썬 버전
    - numpy : 수치해석/통계 관련 작업 지원
    - matplotlib, seaborn : 여러 시각화 도구를 제공
    - Scikit-learn, TensorFlow : 다양한 머신러닝/딥러닝 모델 지원
2. 데이터 분석을 위한 데이터 정제를 쉽게 할 수 있도록 도와준다.
    - 각 자료형의 다양한 기능을 활용하여 데이터 정제
    - 함수/조건/반복을 활용하여 쉽게 데이터 정제

(https://dsstudy.tistory.com/8)- 이 사이트에서 기본적인 코드 학습하였고, 기본적인 그래프 그리는 연습은 colab에서 진행 후 깃허브 계정에 업로드하였음. (공부하는 과정중 테디노트, 나도코딩, 조코딩, 동빈나 등 수많은 유튜브 참조와 구글링 활용하였음)

colab의 가장 큰 장점은 깃허브와의 연동성이라고 할 수 있음. 또한 그래프 구현을 위한 별도 툴 설치가 불필요하다는 점, 상대적으로 작업을 간편하게 할 수 있다는 점 등이 장점이었음.

#코랩과 깃헙 연동시키는 팁, 코랩에서 바로 실습할 수 있어서 매우 편리한 기능

1. Github.com -> github로 주소 변경하기
2. Colab.research.google.com 맨 앞에 추가

프로젝트 과제 요약 분석

.- x축 문자열, y축 시간 데이터가 txt파일에 들어있음.

- txt파일을 만들기 위한 전처리가 깃허브에 있는 metadata.py와 preprocess

- kaggle 데이터셋에는 전처리가 되어있는 txt가 있음.

=> **결국 txt를 그룹화해서 plot하는게 프로젝트의 핵심!**

6번 그래프를 구현하기 어려웠던 이유: 코랩을 이용해서 그래프를 만드는데 Korean 패키지가 실행이 되지 않아서 그래프 plot에 어려움을 겪었음. (Korean 패키지는 Python2 기반으로 제작되어진 것 같습니다.)