기초빅데이터프로그래밍 기말 프로젝트

20180071 오서영

프로그램명: **날씨 정보 프로그램**

2022년 날씨 데이터를 받아 가공한 뒤, 사용자의 입력값에 따라 올해 날씨 정보를 알려주는 프로그램이다. 이때 크롤링을 하기 위한 웹사이트의 복잡도를 고려하여 데이터 측정 기간은 2022년의 날씨 중 관측값이 있는 2022년 1월~2022년 5월 데이터로 한정하였다.

1. 날씨 데이터 크롤링

웹 크롤링을 활용해서 기상청의 2022년 날씨 데이터를 크롤링한다.

(출처: https://www.weather.go.kr/w/obs-climate/land/past-obs/obs-by-day.do)

1. 데이터 전처리

날씨 데이터를 입력 받아 변수에 저장한다. 저장한 날씨 데이터에서 정규식을 활용해 날짜, 요일, 평균기온, 최고기온, 최저기온, 평균운량, 일강수량을 추출한다.

1. OOP

반복문을 통해 '원하는 날씨 정보를 입력하세요.'라는 문구를 출력하고 사용자의 입력값을 받는다.

3-1. 2022년 날씨 정보

사용자가 ‘1’을 입력하면, 2022년의 날씨 정보를 나타낸다. 올해의 최고기온, 올해의 최저기온, 올해의 평균기온, 평균 강수량을 출력한다. 또한 올해의 기온과 강수량 추이를 matplotlib 모듈을 활용해 그래프로 보여준다.`

3-2. 2022년 n월 날씨 정보

사용자가 ‘2’를 입력하면, “몇월의 날씨를 표시할까요?”라는 문구를 출력한다. 사용자가 입력한 해당 월의 최고기온, 최저기온, 평균기온, 평균강수량을 출력한다. 또한 해당 월의 기온과 강수량 추이를 matplotlib 모듈을 활용해 그래프로 보여준다.

3-3. 2022년 n월 n일 날씨 정보

사용자가 ‘3’을 입력하면, “몇월 며칠의 날씨를 표시할까요?”라는 문구를 출력한다. 사용자가 입력한 해당 날짜의 최고기온, 최저기온, 일강수량을 보여준다. 원래 해당 요일의 날씨 정보(rainy, snowy, cloudy, sunny)를 출력하려 하였으나 날씨 선정 기준이 모호하고 강수량의 경우 원본 데이터에 누락값이 있어 이 기능은 생략하였습니다.

1. 에러처리

만약 사용자가 잘못된 날짜를 입력한 경우, “잘못 입력하셨습니다.”라는 에러 문구를 출력한다.

1. 실행시간

하나의 기능을 실행할 때마다 실행시간을 측정하고 화면에 출력한다.

**1. 날씨 데이터 크롤링**

날씨 데이터는 기상청의 과거관측-일별자료를 크롤링하여 추출하였습니다. 단순 추출뿐만 아니라 select나 click처럼 웹페이지 자동화가 필요한 케이스여서 selenium과 beautifulsoup 모듈을 모두 활용하였습니다.

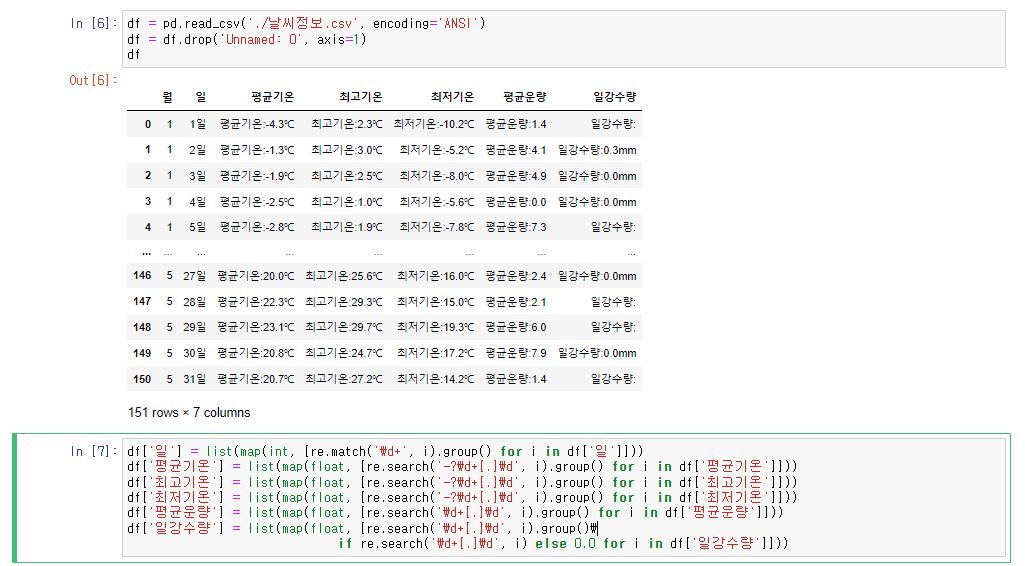


우선 필요한 모듈을 모두 불러주고, 크롬 웹드라이버를 불러와 기상청 웹사이트에 접속합니다. 그리고 weather란 데이터프레임에 '월', '일', '평균기온', '최고기온', '최저기온', '평균운량', '일강수량'을 column으로 하여 크롤링한 데이터를 저장합니다. 이때 각 row는 하루의 데이터가 됩니다.

C:\Users\LOVE.DESKTOP-5CS2DDB\AppData\Local\Microsoft\Windows\INetCache\Content.Word\3.jpg

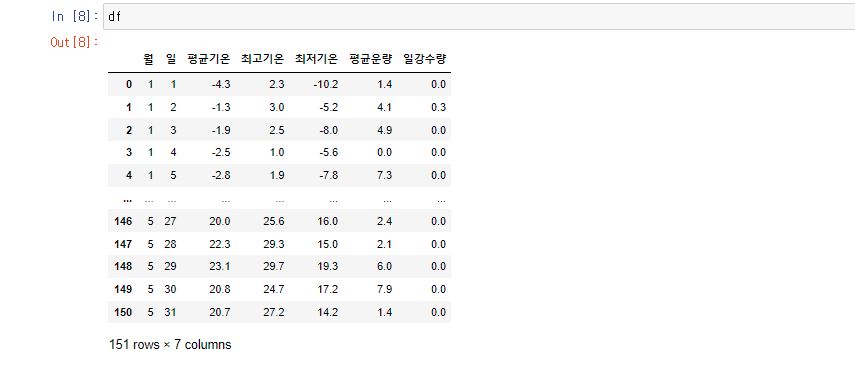
크롤링한 데이터를 '날씨정보.csv'에 to\_csv함수를 활용하여 저장합니다.

**2.** **데이터 전처리**

****

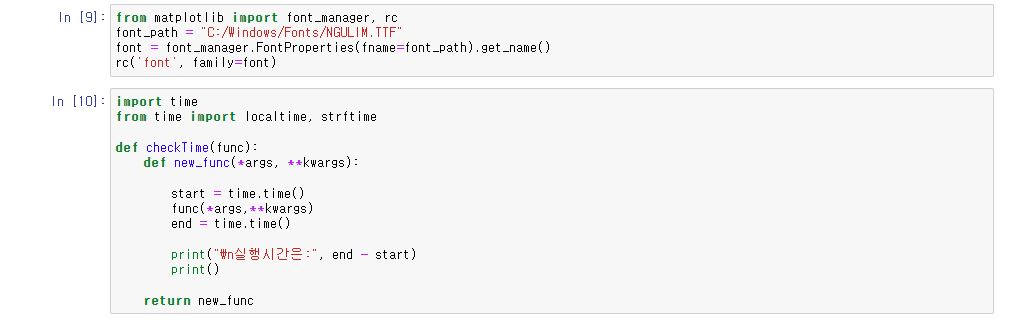
저장한 '날씨정보.csv' 파일을 pandas의 read\_csv함수를 활용하여 df 변수에 읽어옵니다. 이때 인덱스 번호가 'Unnamed: 0' 칼럼에 저장되어 있는데, 불필요한 정보이므로 drop합니다.

df를 확인해보니 한국어 문자와 통계값이 섞여있습니다. 우리가 활용해야 하는 것은 통계값이므로 정규표현식을 활용하여 통계량만 추출합니다. '월'은 정수로 저장되어 있으므로 처리가 따로 필요없습니다. 반면 '일'은 '**(숫자)**일'을 추출해야 하므로 '\d+'를 사용합니다. '평균기온'은 '평균기온**(-숫자.숫자)**℃'를 추출해야 하므로 '-?\d+[.]\d'를 활용하고, '최고기온'과 '최저기온'도 문자열 형식이 동일하므로 똑같이 처리합니다. 마지막으로 평균운량과 일강수량은 음수값이 없고 0.0부터 시작하므로 '\d+[.]\d'으로 처리합니다. 이때 일강수량에서 누락값이 있었는데, 이는 전부 0.0으로 처리하였습니다.

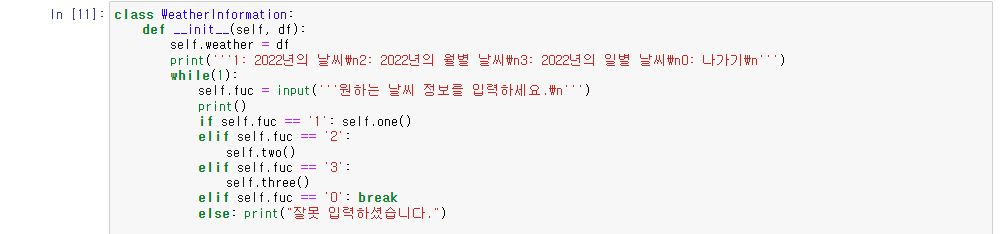


df를 확인해보면 깔끔하게 처리된 것을 확인할 수 있습니다.

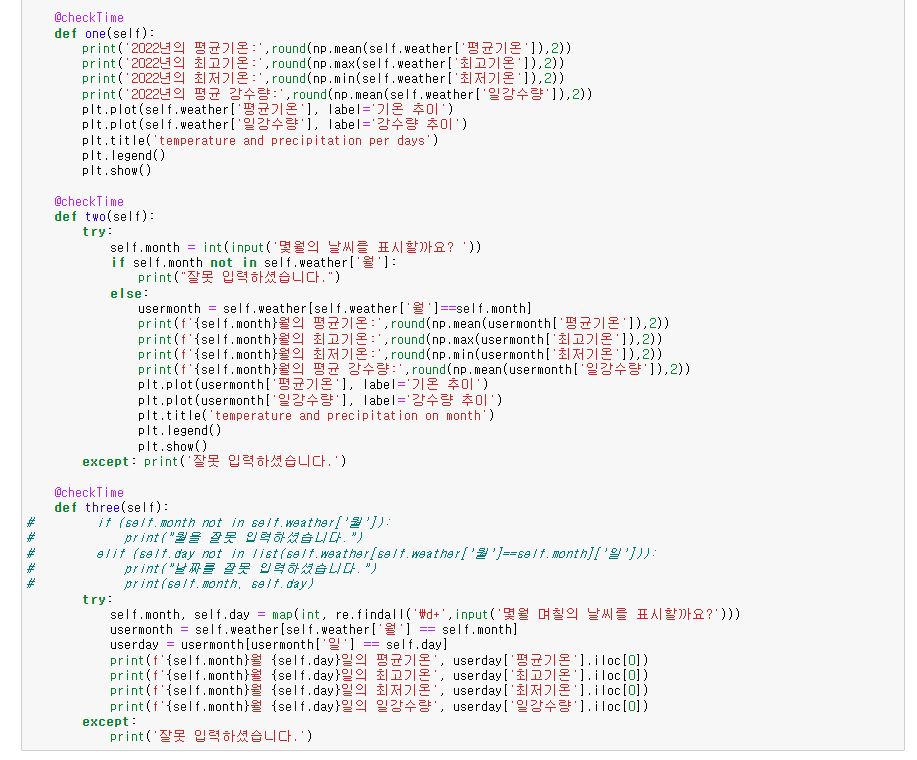
**3. OOP**



그래프를 그릴 때 한글을 활용하기 위해 한글폰트를 불러와줍니다. 그리고 실행시간 측정을 위한 함수를 정의합니다.



날씨 정보를 제공하는 WeatherInformation 클래스를 정의합니다. 날씨 데이터 df를 파라미터로 받아와 self.weather에 저장하고, 반복문을 실행하여 사용자가 원하는 날씨 정보 입력값을 받습니다. 이때 1은 2022년 전체의 날씨 정보, 2는 2022년의 월별 날씨 정보, 3은 2022년의 일별 날씨 정보입니다. 사용자의 입력값에 해당하는 함수를 실행하고, 만약 잘못된 값을 입력한 경우 '잘못 입력하셨습니다.'라는 문구를 출력하고, 혹은 0을 입력한 경우 반복문을 종료합니다.



**one(): 2022년의 날씨 정보를 알려주는 함수**

평균기온, 최고기온, 최저기온, 평균 강수량을 구하기 위해 self.weather 전체 row에 대한 numpy 모듈의 mean, max, min 통계값을 출력합니다. 또한 matplotlib.pyplot의 plot함수를 활용하여 2022년의 평균기온과 일강수량 그래프를 화면에 띄웁니다.

**two(): 2022년의 월별 정보를 알려주는 함수**

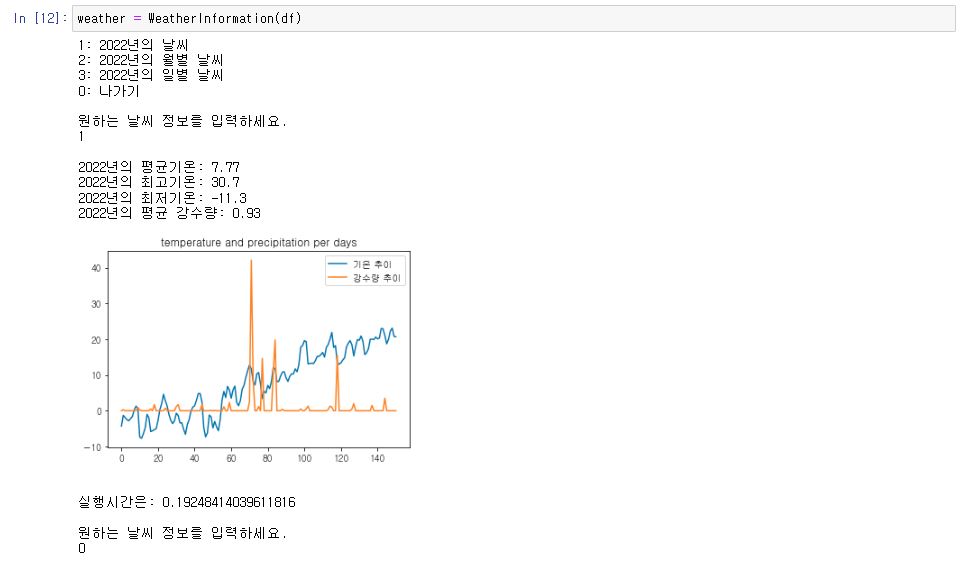
'몇월의 날씨를 표시할까요?'라는 문구를 통해 사용자의 입력값을 받습니다. 만약 입력값이 데이터에 없는 월이면 예외처리를 통해 '잘못 입력하셨습니다.'라는 문구를 출력합니다. 입력값이 정상적인 값이라면 해당 월에 해당하는 row만 추출하여 변수 usermonth에 저장하고, one()에서 사용했던 기능을 동일하게 실행합니다.

**three(): 2022년의 일별 정보를 알려주는 함수**

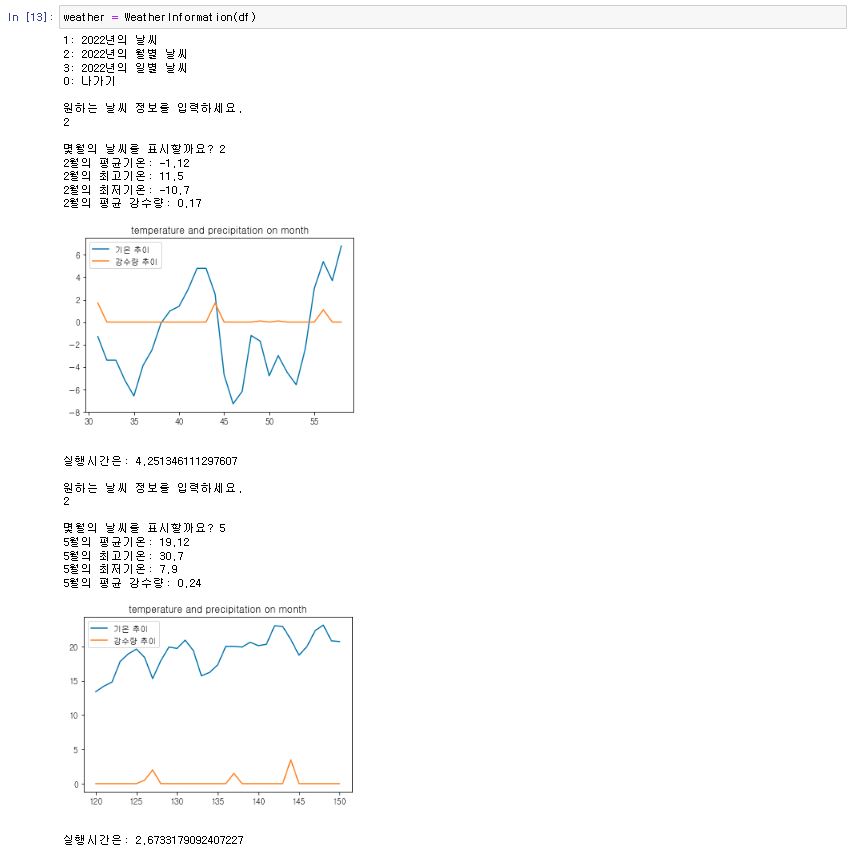
'몇월 며칠의 날씨를 표시할까요?'라는 문구를 통해 사용자의 입력값을 받습니다. 정규표현식을 통해 (월, 일)에 해당하는 정수값을 추출하고 변수 self.month와 self.day에 저장합니다. 만약 입력값이 데이터에 없는 날짜이면 예외처리를 통해 '잘못 입력하셨습니다.'라는 문구를 출력합니다. 입력값이 정상적인 값이라면 해당 날짜에 해당하는 하나의 row를 추출하여 변수 userday에 저장하고, userday의 통계값을 출력합니다.

one(), two(), three()이 실행될 때마다 checkTime()함수가 데코레이터로 실행되어 해당 함수의 실행시간을 출력합니다.

**4. Output**

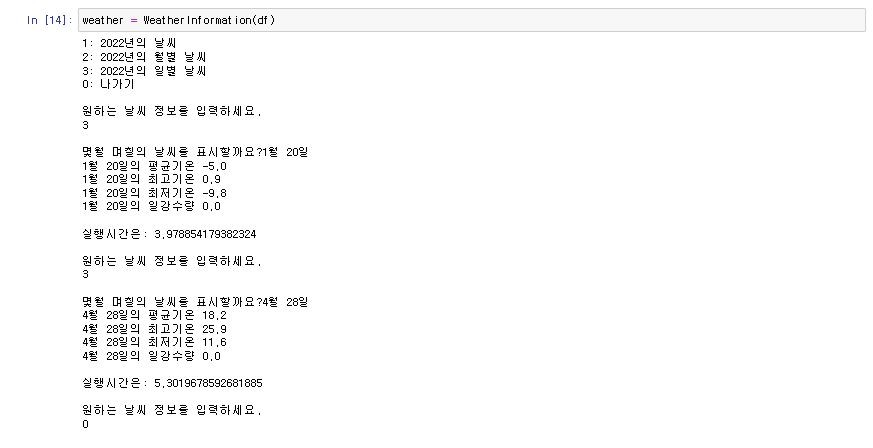


2022년의 날씨 정보입니다. 1을 입력하자, 평균기온, 최고기온, 최저기온, 평균 강수량이 출력되었습니다. 1월부터 5월의 날씨 정보이므로 점차 기온 추이가 올라가는 것을 확인할 수 있습니다.



2022년 N월의 날씨 정보입니다. 2를 입력하면, '몇월의 날씨를 표시할까요?'라는 문구가 출력됩니다. 2를 입력하자, 2월의 통계량을 출력하고 2월 한달 동안의 기온과 강수량 그래프를 띄웁니다. 다시 2를 눌러 실행하고 5를 입력하자, 5월의 통계량을 출력하고 기온과 강수량 그래프를 표시합니다. 2월은 영상과 영하 기온이 반복되는 반면, 5월은 영상에서 점차 기온이 올라가는 추세를 확인할 수 있습니다.

이때 실행시간이 1번에 비해 오래 소요되는 것을 확인할 수 있습니다. 1번은 self.weather 전체 row를 사용하여 통계함수를 처리하는 반면, 2번은 self.weather 내에서 해당 월에 맞는 row만 추출하여 사용합니다. 이러한 차이가 실행시간의 지연을 만든 것으로 추측할 수 있습니다.



2022년 N월 N일의 날씨 정보입니다. 3을 입력하고, 1월 20일이라는 입력값을 주자 1월 20일의 통계량이 출력됩니다. 또 3을 입력하고, 2월 28일이라는 입력값을 주자 2월 28일의 통계량이 출력됩니다. 2번과 마찬가지로 3번도 특정 row를 출력하여 사용하므로 실행시간이 1번에 비해 긴 것을 확인할 수 있습니다.