데이터 사이언스를 위한

파이썬 프로그래밍

텀 프로젝트 계획서

2020.06.04



산업시스템공학과

2017112526

김민지

1. 주제선정

네이버 날씨에 나오는 지역별 날씨 정보를 이용하여, 시계열 데이터 및 정보를 제공한다.

1. 계획

* 데이터 정보 및 수집 방법

1. 지역 : 서울특별시 (중구), 제천시 (고암동), 대전광역시(대덕구), 대구광역시 (중구), 부산광역시 (중구)로 설정
2. 날씨 정보 : 매일 오후 2시경 현재 날씨(흐림 등) ,현재온도(21℃), 미세먼지 정도(보통/낮음 등) , 오존 정도(매우 높음/낮음 등)
3. 정해진 시간(오후 2시)에 매일 데이터를 웹 크롤링(Beautiful Soup)를 이용하여 수집한다.
4. 수집한 데이터를 Maria DB에 저장한다.
5. 사용기술

* 웹 크롤링 : 웹에서 유용한 정보를 찾아 특정 데이터베이스로 수집해오는 기술
* Package : BeautifulSoup 사용
* 지역별 날씨정보를 크롤링 기술을 사용하여, 자료를 수집한다.
* 데이터 시리즈 및 프레임 다루기 및 데이터베이스
* Package : Pandas / pymysql 사용
* 수집한 데이터를 2차원 자료구조인 데이터 프레임을 이용하여 정리한다.
* Column은 날짜, 현재 날씨, 현재 온도, 미세먼지 정도, 오존 정도로 설정한다.
* Group by 함수를 사용하여 지역별로 혹은 날짜별로 정보를 얻을 수 있도록 한다.
* 해당 내용을 pymysql패키지를 이용해 데이터베이스에 저장한다.
* 데이터 시각화
* Package : matplotlib.pyplot, matplotlib 사용
* 해당 패키지를 이용하여, 그래프를 그려 시각화 된 자료를 제공한다.
* 날짜 별 각 지역의 현재 온도 시계열 그래프 / 한 지역의 현재 온도변화 시계열 그래프

1. 입력 데이터 출처

* 네이버 날씨 웹사이트 : <https://weather.naver.com/>
* 해당 웹사이트 설정 이유
* 원하는 지역을 직접 골라 원하는 정보를 데이터 베이스를 만들 수 있음

1. 출력 데이터 양식 및 사용할 프로그램 설명

예상 데이터베이스 형식

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 지역 | 날짜 | 현재 날씨 | 현재 온도 | 미세먼지 정도 | 오존 정도 |
| 서울 | 20200603 | 맑음 | 28 | 보통 | 보통 |
| 제천 | 20200603 | 흐림 | 24 | 좋음 | 보통 |
| 대전 | 20200603 | 흐림 | 24 | 보통 | 보통 |
| 대구 | 20200603 | 맑음 | 29 | 나쁨 | 나쁨 |
| 부산 | 20200603 | 흐림 | 26 | 좋음 | 나쁨 |
| 서울 | 20200604 | 맑음 | 28 | 보통 | 보통 |
| 제천 | 20200604 | 흐림 | 24 | 좋음 | 보통 |
| 대전 | 20200604 | 흐림 | 24 | 보통 | 보통 |
| 대구 | 20200604 | 맑음 | 29 | 나쁨 | 나쁨 |
| 부산 | 20200604 | 흐림 | 26 | 좋음 | 나쁨 |

메뉴 형식으로 해당 항목들을 입력 받아서 다음과 같은 행동을 취한다.

1. 원하는 지역, 날짜를 입력하면, 날씨 정보 모두 제공

input으로 지역, 날짜를 입력 받으면 그것에 해당하는 현재 날씨, 현재온도, 미세먼지 정도, 오존 정도를 모두 제공한다.

1. 원하는 지역, 날짜, 날씨 정보를 입력하면, 원하는 정보 제공

input으로 지역, 날짜와 날씨 정보를 모두 입력 받아 원하는 정보를 제공한다.

1. 전체 지역의 시계열 현재온도 시계열 그래프

지역별로 Groupby 하여서, 각 지역마다 날짜별로 현재온도의 변화를 그래프로 시각화 자료를 제공한다.

1. 원하는 지역 입력 시 해당지역의 현재온도 시계열 그래프

지역별로 Groupby 하여, 원하는 지역의 날짜별로 현재온도의 변화를 그래프로 시각화 자료를 제공한다.

1. 날짜 별, 현재 온도가 최고 높은 지역 및 온도 제공

날짜로 Groupby해서, 날짜별로 현재온도가 가장 높은 지역과 그 온도를 제공한다.

1. 과거부터 현재까지 데이터를 보았을 때, 미세먼지/오존 정도가 가장 나쁜 지역

지역별로 Groupby미세먼지와 오존정도 (좋음/보통/나쁨)에 가중치를 부여해, 현재까지 가장 미세먼지/오존 정도가 나쁜 지역을 뽑아 제공한다.