

빅데이터 수집 시스템 구성

김은표

목차

- ◉ 네이버 실시간 검색어 수집
- ◉ 수집 데이터 파일로 저장
- ◉ 전국 날씨 데이터 수집
- ◉ 수집 데이터 파일로 저장
- ◉ Crontab 설정

네이버 실시간 검색어 수집

```
12 # 리눅스용 Chrome 브라우저 설치
13 # 크롬 가상 웹브라우저 실행 (headless 모드)
14 chrome_option = webdriver.ChromeOptions()
15 chrome_option.add_argument('--headless')
16 chrome_option.add_argument('--no-sandbox')
17 chrome_option.add_argument('--disable-dev-shm-usage')
18 browser = webdriver.Chrome('./chromedriver', options=chrome_option)
19 browser.implicitly_wait(3)
20
21 # 네이버 데이터랩 이용
22 browser.get('https://datalab.naver.com/keyword/realtimeList.naver')
23 browser.implicitly_wait(3)
24
25 # 네이버 실검 1 ~ 10까지 파싱
26 item_boxs = browser.find_elements_by_css_selector('#content .selection_area .field')
27
28 # 디렉터리 생성
29 dir = "./naver/{:%Y-%m-%d}".format(datetime.now())
30
31 if not os.path.exists(dir):
32     os.makedirs(dir)
33
34 # 파일 저장
35 fname = "{:%y-%m-%d-%H-%M}.txt".format(datetime.now())
36 file = open(dir+'/'+fname, mode='w', encoding='utf8')
37 file.write('순 위 , 제 목 , 날 짜 \n')
38
39 for item_box in item_boxs:
40     file.write('%s, ' % item_box.find_element_by_css_selector('.item_num').text)
41     file.write('%s, ' % item_box.find_element_by_css_selector('.item_title').text)
42     file.write('%s\n' % "{:%y%m%d%H%M%S}".format(datetime.now()))
43
```

파이썬을 이용하여 네이버 실시간 검색어 수집 소스코드 작성하여 리눅스와 연동
네이버 실검 파싱하여 디렉터리를 생성하고 수집된 데이터를 저장하게 설정

수집 데이터 파일로 저장

```
2_7_naver.py × 20-12-31-10-57.txt ×
1 순위, 제목, 날짜
2 1, 2021년 새해인사말, 201231105707
3 2, 곧대인턴, 201231105707
4 3, 2020 mbc 연기대상, 201231105707
5 4, 새해 인사말, 201231105707
6 5, 유명민, 201231105707
7 6, 12월 31일 주식, 201231105707
8 7, 박해진, 201231105707
9 8, 이재웅, 201231105707
10 9, 연말 인사말, 201231105707
11 10, 12월 31일 주식시장, 201231105707
12
```

급상승검색어 관심사/연령대별 나만의 차트를 설정할 수 있어요. ✕

검색어 설정 ①

이슈별 묶어보기

이벤트 · 할인

시사

엔터

스포츠

1~10위 11~20위

- 1 2021년 새해인사말 ∞ 새해 인사말
- 2 2021년 새해인사 ∞ 새해인사, 연말 인사말, ...
- 3 인스타베스트나인
- 4 유명민
- 5 12월 31일 주식
- 6 주식시장
- 7 이재웅
- 8 12월 31일 주식시장
- 9 파랑새노인건강센터
- 10 종무식

10대	20대	30대
40대	50대	전체연령

전국 날씨 데이터 수집

```
5 import os
6 import requests as req
7 from bs4 import BeautifulSoup as bs
8 from datetime import datetime
9 from selenium import webdriver
10
11 chrome_option = webdriver.ChromeOptions()
12 chrome_option.add_argument('--headless')
13 chrome_option.add_argument('--no-sandbox')
14 chrome_option.add_argument('--disable-dev-shm-usage')
15 browser = webdriver.Chrome('./chromedriver', options=chrome_opti
16 browser.implicitly_wait(3)
17
18 browser.get('https://www.weather.go.kr/w/weather/now.do')
19 browser.implicitly_wait(3)
20
21 trs = browser.find_elements_by_css_selector('#sfc-city-weather >
22
23 # 디렉터리 생성
24 dir = "/home/bigdata/weather/{:%Y-%m-%d}".format(datetime.now())
25
26 if not os.path.exists(dir):
27     os.makedirs(dir)
28
29
30 # 파일 저장
31 fname = "{:%y-%m-%d-%H-%M.txt}".format(datetime.now())
32 file = open(dir+'/'+fname, mode='w', encoding='utf-8')
33
34 file.write('지점,현재일기,시정,운량,증하운량,현재기온,이슬점온도
35
```

파이썬을 이용하여 기상청 전국 날씨 데이터 수집 소스코드 작성하여 리눅스와 연동
기상청 전국 날씨 파싱하여 디렉터리를 생성하고 수집된 데이터를 저장하게 설정

지점, 현재일기, 시정, 운량, 중하운량, 현재기온, 이슬점온도, 일강수, 적설, 습도, 풍향, 풍속, 해면기압

강릉, ,20 이상, , , -3.6, -19.5, -8.6, , , 28, 서남서, 13.7, 1015.1
강진군, ,12.5, , , -2.4, -7.7, -7.3, , , 67, 북서, 14.4, 1022.3
강화, 맑음, 20 이상, 0, 0, -9.0, -18.5, -14.4, , , 46, 서북서, 11.2, 1023.6
거제, ,20 이상, , , -1.2, -15.4, -6.3, 1.3, , 33, 서북서, 16.9, 1018.4
거창, ,20 이상, , , -5.6, -15.9, -12.4, 0.0, , 44, 북서, 20.5, 1018.9
경주시, ,20 이상, , , -4.1, -18.7, -10.5, 0.6, , 31, 북서, 20.5, 1017.5
고산, ,20 이상, , , 0.9, -2.1, -8.6, 0.2, , 80, 북서, 82.8, 1021.0
고창, ,3.3, , , -3.6, -10.0, -11.7, 3.1, 2.0, 61, 서북서, 34.9, 1022.5
고창군, ,1.9, , , -6.9, -9.4, -10.8, 0.0, 9.1, 82, 북북서, 7.9, 1023.0
고흥, ,20 이상, , , -3.2, -10.3, -8.7, 0.0, , 58, 북북서, 16.6, 1020.7
광양시, ,20 이상, , , -3.5, -12.4, -9.1, 1.4, , 50, 서북서, 16.9, 1020.5
광주, 약한 눈 연속적, 1.7, 10, 9, -5.0, -6.9, -7.3, 4.0, 16.3, 86, 북북동, 5.0, 1022.4
구미, ,20 이상, , , -4.8, -17.9, -10.2, 0.0, , 35, 북북서, 14.4, 1020.0
군산, , , , , -5.4, -11.1, -12.1, 3.5, 4.3, 64, 북북서, 20.2, 1023.2
금산, ,19.2, , , -7.2, -13.2, -12.8, 0.4, 3.3, 62, 북서, 13.0, 1022.5
김해시, ,20 이상, , , -2.2, -18.7, -7.9, 0.7, , 27, 북서, 19.4, 1018.1
남원, ,2.9, , , -7.0, -9.7, -10.8, 2.2, 5.3, 81, 북북동, 7.6, 1022.6
남해, ,20 이상, , , -2.6, -12.6, -8.3, 0.6, , 46, 북서, 18.7, 1019.7
대관령, ,20 이상, , , -13.1, -20.9, -20.7, , 1.0, 52, 서, 15.5, 1017.8
대구, 맑음, 20 이상, 1, 1, -4.5, -17.0, -12.2, 0.2, , 37, 북서, 28.4, 1018.9
대전, 약한 눈 단속적, 20 이상, 4, 4, -6.5, -14.9, -12.5, 0.0, 1.2, 51, 북북동, 15.1, 1022.7
동두천, ,19.4, , , -7.8, -19.6, -13.0, , , 38, 서, 11.2, 1022.8
동해, ,20 이상, , , -3.6, -23.3, -8.2, , , 20, 북서, 12.2, 1014.2
목포, 약한 소낙눈, 13.2, 8, 8, -4.0, -6.7, -12.9, 0.5, 0.0, 81, 북북서, 41.8, 1022.8
문경, ,18.4, , , -7.3, -20.1, -14.2, , , 35, 서북서, 18.7, 1019.1
밀양, 맑음, 20 이상, 1, 0, -2.2, -18.7, -7.3, 0.0, 0.7, 27, 북북서, 15.8, 1017.5
백령도, 약한 눈 연속적, 5.1, 9, 9, -9.0, -12.3, -19.7, 0.6, 1.1, 77, 북북서, 42.8, 1024.3
보령, ,18.9, , , -6.1, -12.2, -10.5, 1.6, 2.5, 62, 북, 9.7, 1022.2
보성군, ,15.3, , , -2.7, -6.8, -10.7, 0.1, , 73, 북북서, 36.4, 1020.7
보은, ,20 이상, , , -7.9, -13.7, -13.5, 1.2, 2.7, 63, 서북서, 12.6, 1022.1
봉화, ,20 이상, , , -7.9, -21.4, -13.0, , , 33, 서남서, 10.8, 1017.7
부산, 맑음, 20 이상, 1, 1, -1.7, -18.2, -7.5, 1.2, , 27, 서북서, 20.5, 1016.5
부안, ,1.2, , , -6.4, -7.6, -6.4, 13.4, 23.4, 91, 북북동, 4.3, 1023.2
부여, ,10.1, , , -6.1, -12.0, -11.0, 0.1, , 50, 북서, 11.0, 1023.6

Crontab 설정

```
root@Bigdata101:/home/bigdata/weather
```

```
* * * * * python3 /root/2_7_naver.py  
10 * * * * python3 /root/2_8_weather.py  
10 * * * * python3 /root/2_9_weather_to_db.py
```

crontab -e 를 이용하여 크론탭을 편집할 수 있다.

각 별 위치에 따라 주기를 다르게 설정 할 수 있습니다. 순서대로 분-시간-일-월-요일 순입니다. 그리고 괄호 안의 숫자 범위 내로 별 대신 입력 할 수 있습니다.

요일에서 0과 7은 일요일입니다. 1부터 월요일이고 6이 토요일입니다.

첫째줄 네이버 실시간 검색 크론탭 설정은 (* * * * *)으로 매분마다 자동 실행된다.

둘째줄 날씨 데이터 크론탭 설정은 (10 * * * *)으로 매시간 10분에 자동실행이 된다.

마지막 줄 크론탭 설정은 (10 * * * *)으로 설정되어 있기에 매시간 10분에 자동 된다.