

# 빅데이터 수집 시스템

빅데이터 수집 시스템 구성 및  
빅데이터 수집/변환 모듈 개발

# 수행 순서

## 1. 빅데이터 수집 시스템 구성

- ① 네이버 실시간 검색어 수집
- ② 1분 마다 한번 Crawler 실행
- ③ 수집 데이터 파일로 저장

## 2. 빅데이터 수집/변환 모듈 개발

- ① 전국 날씨 데이터 수집
- ② 1시간 마다 한번 Crawler 실행
- ③ 수집 데이터 파일로 저장

# 1. 빅데이터 수집 시스템 구성 - 네이버 실시간 검색어

- 네이버 데이터랩 급상승 검색어 (<https://datalab.naver.com/keyword/realtimeList.naver>) 웹에서 HTML 코드를 확인해서 찾고자하는 문자열, 크롤링할 문자들의 태그와 클래스를 복사한다.

NAVER DataLab.

데이터랩 홈 급상승검색어 검색어트렌드 쇼파인사이트 지역통계 댓글통계

### 급상승 검색어

검색 횟수가 급상승한 검색어의 순위를 다양한 옵션을 통해 자세히 제공함

집계 주기 < 2020.12.31. (목) > < 11:38 >

상세 옵션

- 이슈별 묶어보기
- 이벤트 · 할인
- 시사
- 엔터
- 스포츠

과거 데이터 조회하기

2017.03.29. ~ 2018.10.10. >

2018.10.10. ~ 2019.11.28. >

2019.11.28. ~ 2020.01.15. >

- 1 2021년 새해 인사말
- 2 새해 인사말
- 3 제이미
- 4 곧대인턴
- 5 2020 mbc 연기대상
- 6 제이미 박지민 ⇄ 박지민
- 7 유미의 세로돌
- 8 유명민
- 9 서정희나이

span.item\_title 125.5 x 15

```
<div id="content" class="content">
  <div class="section_keyword">
    <div class="com_title title_keyword">...</div>
    <div class="selection_area">
      <div class="selection_header">...</div>
      <!-- 상세옵션 및 대표검색어, 유사검색어 ranking -->
      <div class="selection_content">
        <div class="field_option">...</div>
        <div class="field_list">
          <div class="ranking_box">
            <div class="list_group">
              ::before
              <ul class="ranking_list">
                <li class="ranking_item">
                  <div class="item_box">
                    <span class="item_num">1</span>
                    <span class="item_title_wrap">
                      <span class="item_title">2021년 새해인사말</span> == $0
                    </span>
                  </div>
                </li>
                <li class="ranking_item">...</li>
                <li class="ranking_item">...</li>
                <li class="ranking_item">...</li>
                <li class="ranking_item">...</li>
                <li class="ranking_item">...</li>
                <li class="ranking_item">...</li>
                <li class="ranking_item">...</li>
                <li class="ranking_item">...</li>
                <li class="ranking_item">...</li>
                <li class="ranking_item">...</li>
              </ul>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
```

# 1. 빅데이터 수집 시스템 구성 – 네이버 실시간 검색어

- 먼저 Html을 쉽게 파싱해 줄 수 있는 BeautifulSoup의 bs4 모듈과 http요청을 할 수 있는 requests 모듈을 설치해야한다.
- 또한, Selenium 모듈은 주로 웹앱을 테스트하는데 이용하는 프레임워크다. webdriver라는 API를 통해 운영체제에 설치된 Chrome등의 브라우저를 제어하게 된다.
- 크롬을 사용하려면 로컬에 크롬이 설치되어있어야 한다. 다음 링크를 클릭하여 리눅스용 크롬드라이버를 다운받아 리눅스로 파일을 옮긴다. (링크: <https://sites.google.com/a/chromium.org/chromedriver/downloads>)

```
[root@Bigdata101 2020-12-31]# cd ~
[root@Bigdata101 ~]# ll
합 계 10960
-rw-r--r--  1 root root      1591 12월  31 11:29 2_7_naver.py
-rw-----  1 root root      1389 12월  29 15:16 anaconda-ks.cfg
-rwxr-xr-x  1 root root 11214464 12월  29 17:02 chromedriver
[root@Bigdata101 ~]#
```

# 1. 빅데이터 수집 시스템 구성 - 네이버 실시간 검색어

- 다음과 같이 python을 이용해 네이버 실시간 검색어 크롤링할 수 있는 코드를 작성한다.

```
6 import os
7 import requests as req
8 from bs4 import BeautifulSoup as bs
9 from selenium import webdriver
10 from datetime import datetime
11
12 # 크롬 가상 웹브라우저 실행 (headless 모드)
13 chrome_option = webdriver.ChromeOptions()
14 chrome_option.add_argument('--headless')
15 chrome_option.add_argument('--no-sandbox')
16 chrome_option.add_argument('--disable-dev-shm-usage')
17 browser = webdriver.Chrome('./chromedriver', options=chrome_option)
18 browser.implicitly_wait(3)
19
20 # 네이버 데이터랩 이동
21 browser.get('https://datalab.naver.com/keyword/realtimeList.naver')
22 browser.implicitly_wait(3)
23
24 # 네이버 실검 1 ~ 10까지 파싱 - <div></div>....10개
25 item_boxes = browser.find_elements_by_css_selector('#content .selection_area .field_list ul:nth-child(1) > li > '
26                                                    '.item_box')
27
28 # 디렉토리 생성
29 dir = "/home/bigdata/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_boxes:
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
44 file.close()
45
46 browser.close()
47 print('수 집 완 료 ...')
```

# 1. 빅데이터 수집 시스템 구성 – 네이버 실시간 검색어

- 1분 마다 한번 Crawler를 실행하기 위해서는 리눅스 작업 스케줄러인 crontab을 이용한다.
- 실행방법: #crontab -e
- crontab -e 를 이용해 다음과 같이 명령어를 사용하여 프로세스를 실행한다.

```
* * * * * python3 /root/2_7_naver.py
```

- 해당 날짜로 생성된 (2020-12-31) 디렉토리(# cd /home/bigdata/naver)를 확인하면 1분 마다 수집 데이터 파일로 저장된 것을 확인할 수 있다.

```
[root@Bigdata101 naver]# ll
합 계 0
drwxr-xr-x 2 root root 136 12월 31 11:34 2020-12-31
[root@Bigdata101 naver]# cd 2020-12-31/
[root@Bigdata101 2020-12-31]# ll
합 계 20
-rw-r--r-- 1 root root 329 12월 31 11:30 20-12-31-11-30.txt
-rw-r--r-- 1 root root 328 12월 31 11:31 20-12-31-11-31.txt
-rw-r--r-- 1 root root 328 12월 31 11:32 20-12-31-11-32.txt
-rw-r--r-- 1 root root 334 12월 31 11:33 20-12-31-11-33.txt
-rw-r--r-- 1 root root 334 12월 31 11:34 20-12-31-11-34.txt
[root@Bigdata101 2020-12-31]#
```

# 1. 빅데이터 수집 시스템 구성 – 네이버 실시간 검색어

- 아래는 수집된 네이버 실시간 검색어 1위부터 10위를 확인할 수 있다.

```
1 순위, 제목, 날짜
2 1, 2021년 새해 인사말, 201231113002
3 2, 2021년 새해 인사, 201231113002
4 3, 새해 인사말, 201231113002
5 4, 곧대인턴, 201231113003
6 5, 제이미, 201231113003
7 6, 2020 mbc 연기대상, 201231113003
8 7, 유미의세포들, 201231113003
9 8, 박해진, 201231113003
10 9, 유영민, 201231113003
11 10, 이재용, 201231113003
```

## 2. 빅데이터 수집 시스템 구성 - 전국 날씨 데이터

- 기상청의 현재 전국 날씨 데이터 (<https://www.weather.go.kr/w/weather/now.do>) 를 웹에서 HTML 코드를 확인해서 찾고자하는 문자열, 크롤링할 문자들의 태그와 클래스를 복사한다.

The image shows a screenshot of the Korean Meteorological Administration's website (www.weather.go.kr) displaying current weather information. The left sidebar contains navigation links: 날씨 (Weather), 바다 (Sea), 영상·일기도 (Video/Map), 태풍 (Typhoon), 지진·화산 (Earthquake/Volcano), and 소식·지식 (News/Knowledge). The main content area is titled '현재날씨' (Current Weather) and features a table with weather data. A tooltip is visible over the '강릉' (Gangneung) link, showing its attributes: a.link, 30.27 x 21, Color #000000, Font 17px ns, sans-serif, and Accessibility details (Contrast, Name, Role, Keyboard-focusable). The right side of the image shows the DevTools 'Elements' panel, displaying the HTML structure of the weather data table. The table is a single-column table with a caption '기상실황표로 지점, 날씨, 기온, 강수, 바람, 기압등을 안내한 표입니다.' (Table providing information on location, weather, temperature, precipitation, wind, and pressure using the weather real-time table). The table body contains a row for Gangneung with a link to view the city's weather for 2020.12.31.11:00.

현재날씨		
<table border="1"><thead><tr><th>날씨</th></tr></thead><tbody><tr><td>강릉</td></tr></tbody></table>	날씨	강릉
날씨		
강릉		



## 2. 빅데이터 수집 시스템 구성 - 전국 날씨 데이터

- 다음과 같이 python을 이용해 전국 현재 날씨 데이터의 웹 크롤링(web crawling) 코드를 작성한다.

```
import os
import requests as req
from bs4 import BeautifulSoup as bs
from datetime import datetime
from selenium import webdriver

# 크롬 가상 웹 브라우저 실행 (headless 모드)
chrome_option = webdriver.ChromeOptions()
chrome_option.add_argument('--headless')
chrome_option.add_argument('--no-sandbox')
chrome_option.add_argument('--disable-dev-shm-usage')
browser = webdriver.Chrome('./chromedriver', options=chrome_option)
browser.implicitly_wait(3)

browser.get('https://www.weather.go.kr/w/weather/now.do')
browser.implicitly_wait(3)

trs = browser.find_elements_by_css_selector(
    '#sfc-city-weather > div.cont-box02 > div > div.cont02 > div > table > tbody > tr')

# 디렉터리 생성
dir = "/home/bigdata/weather/({:%Y-%m-%d}).format(datetime.now())

if not os.path.exists(dir):
    os.makedirs(dir)

# 파일 저장
fname = "{:%Y-%m-%d-%H-%M.txt}".format(datetime.now())
file = open(dir+'/' + fname, mode='w', encoding='utf-8')

file.write('지점,현재일기,시점,온도,증하온도,현재기온,이슬점온도,체감온도,일강수,적설,습도,풍향,풍속,해면기압\n')

8 for tr in trs:
9     v1 = tr.find_element_by_css_selector('td:nth-child(1) > a').text
10    v2 = tr.find_element_by_css_selector('td:nth-child(2)').text
11    v3 = tr.find_element_by_css_selector('td:nth-child(3)').text
12    v4 = tr.find_element_by_css_selector('td:nth-child(4)').text
13    v5 = tr.find_element_by_css_selector('td:nth-child(5)').text
14    v6 = tr.find_element_by_css_selector('td:nth-child(6)').text
15    v7 = tr.find_element_by_css_selector('td:nth-child(7)').text
16    v8 = tr.find_element_by_css_selector('td:nth-child(8)').text
17    v9 = tr.find_element_by_css_selector('td:nth-child(9)').text
18    v10 = tr.find_element_by_css_selector('td:nth-child(10)').text
19    v11 = tr.find_element_by_css_selector('td:nth-child(11)').text
20    v12 = tr.find_element_by_css_selector('td:nth-child(12)').text
21    v13 = tr.find_element_by_css_selector('td:nth-child(13)').text
22    v14 = tr.find_element_by_css_selector('td:nth-child(14)').text

3
4     file.write(
5         '%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s,%s\n' % (v1, v2, v3, v4, v5, v6, v7, v8, v9, v10, v11, v12, v13, v14))
6
7 # 파일 닫기
8 file.close()
9
10 # 브라우저 종료
11 browser.close()
12 browser.quit()
13
14 print('날씨 데이터 수집 완료...')
```

## 2. 빅데이터 수집 시스템 구성 – 전국 날씨 데이터

- 1시간 마다 한번 Crawler를 실행하기 위해서는 리눅스 작업 스케줄러인 crontab을 이용한다.
- 실행방법: #crontab -e
- crontab -e 를 이용해 다음과 같이 명령어를 사용하여 프로세스를 실행한다.

```
0 * * * * python3 /root/2_8_weather.py
```

- 해당 날짜로 생성된 (2020-12-31) 디렉토리((# cd /home/bigdata/weather) 를 확인하면 1시간 마다 수집 데이터 파일로 저장된 것을 확인할 수 있다.

```
[root@Bigdata101 2020-12-31]# ll
합 계 8
-rw-r--r-- 1 root root 140 12월 31 12:56 20-12-31-12-00.txt
-rw-r--r-- 1 root root 140 12월 31 13:01 20-12-31-13-00.txt
[root@Bigdata101 2020-12-31]#
```

## 2. 빅데이터 수집 시스템 구성 - 전국 날씨 데이터

- 아래는 수집된 전국 현재 날씨 데이터를 확인할 수 있다.

```
1 |지점,현재일기,시점,운량,증하운량,현재기온,이슬점온도,체감온도,일강수,적설,습도,풍향,풍속,해면기압
2 |강릉,20 이상,-2.9,-20.7,-8.2,,24,남남서,16.2,1021.0
3 |강진군,20 이상,,0.1,-5.5,-3.7,,2.3,66,북서,11.9,1026.8
4 |강화,20 이상,,7.0,-18.0,-11.1,,41,동남동,8.3,1026.9
5 |거제,20 이상,,-0.4,-14.0,-3.1,,35,서북서,7.6,1025.4
6 |거창,20 이상,,-1.8,-12.7,-5.4,,43,북북서,9.7,1024.7
7 |경주시,20 이상,,-3.7,-15.6,-9.8,,39,북서,19.4,1024.4
8 |고산,20 이상,,3.4,-2.5,-3.6,0.7,,65,북북서,52.2,1026.7
9 |고창,20 이상,,-4.0,-8.3,-4.0,0.3,7.4,72,동,1.8,1027.3
10 |고창군,20 이상,,-5.6,-10.0,-5.6,0.0,14.3,71,서북서,3.6,1027.4
11 |고흥,20 이상,,-1.1,-8.5,-4.9,,57,북서,11.2,1025.9
12 |광양시,,14.2,,0.3,-10.2,-5.4,,47,서,18.7,1025.4
13 |광주,맑음,16.4,0.0,-3.6,-9.4,-3.6,,10.6,64,남서,2.9,1027.2
14 |구미,20 이상,,-2.9,-15.2,-6.5,,38,북북서,9.0,1026.4
15 |군산,,,-4.8,-11.3,-8.1,0.0,2.5,60,북북서,7.2,1027.8
16 |금산,,19.1,,-6.5,-13.8,-6.5,,2.4,56,정온,1.1,1027.8
17 |김해시,20 이상,,-2.5,-15.2,-7.1,,37,서북서,13.3,1025.4
18 |남원,20 이상,,-3.5,-10.1,-6.3,,3.0,60,북북서,6.5,1027.6
19 |남해,20 이상,,1.1,-9.8,-1.8,,44,북서,9.4,1025.1
20 |대관령,20 이상,,-12.3,-22.0,-23.2,,44,서,34.6,1022.7
21 |대구,맑음,20 이상,0.0,-3.4,-13.9,-9.5,,44,서북서,19.8,1025.9
22 |대전,맑음,20 이상,0.0,-5.3,-13.6,-8.8,,1.2,52,남,7.6,1027.7
23 |동두천,맑음,20 이상,0.0,-7.0,-18.9,-7.0,,38,남남서,4.0,1026.9
24 |동해,20 이상,,-3.6,-22.8,-8.7,,21,서북서,14.4,1020.9
25 |목포,약한소낙눈,16.0,8.8,-2.9,-7.6,-4.9,0.7,3.5,70,남남동,5.0,1027.5
26 |문경,20 이상,,-5.6,-17.0,-9.6,,40,서북서,8.6,1025.2
27 |밀양,20 이상,,-1.2,-15.0,-4.2,,34,북서,8.3,1024.4
28 |백령도,약한눈연속적,1.1,10.8,-5.6,-6.5,-10.8,2.8,5.2,93,동,13.0,1025.1
29 |보령,,18.7,,-3.9,-11.1,-3.9,0.0,1.0,57,남남동,4.7,1026.1
30 |보성군,20 이상,,0.5,-6.6,-4.0,,59,북서,15.8,1025.7
31 |보은,20 이상,,-7.3,-15.9,-7.3,,1.9,50,남,4.3,1027.2
32 |봉화,20 이상,,-6.9,-19.8,-13.4,,35,북북서,17.3,1024.0
33 |부산,맑음,20 이상,0.0,-2.4,-15.1,-7.9,,37,서,17.6,1023.3
34 |부안,20 이상,,-5.3,-11.6,-5.3,1.0,20.2,61,동남동,4.7,1027.7
35 |부여,,19.8,,-5.1,-12.5,-5.1,0.0,0.1,56,남서,3.6,1028.2
36 |강릉,맑음,20 이상,0.0,-3.5,-23.2,-9.6,,20,서,19.4,1019.8
37 |북창원,20 이상,,-0.6,-12.5,-2.8,,40,서북서,6.1,1024.9
38 |북춘천,맑음,20 이상,0.0,-8.5,-20.6,-11.3,,37,남,5.0,1026.4
39 |산청,20 이상,,-1.0,-11.4,-5.8,,45,서북서,15.8,1025.4
40 |상주,20 이상,,-5.4,-16.6,-9.3,,41,서,8.6,1026.4
41 |서귀포,약한소낙눈,14.7,9.9,2.9,-0.3,2.9,0.7,1.0,79,서,3.2,1026.9
42 |서산,,18.3,,-5.1,-11.4,-5.1,0.0,,61,서,3.2,1026.7
43 |서울,맑음,20 이상,0.0,-7.2,-17.1,-10.7,,45,서북서,6.8,1026.3
```