SW멘토링 프로그램

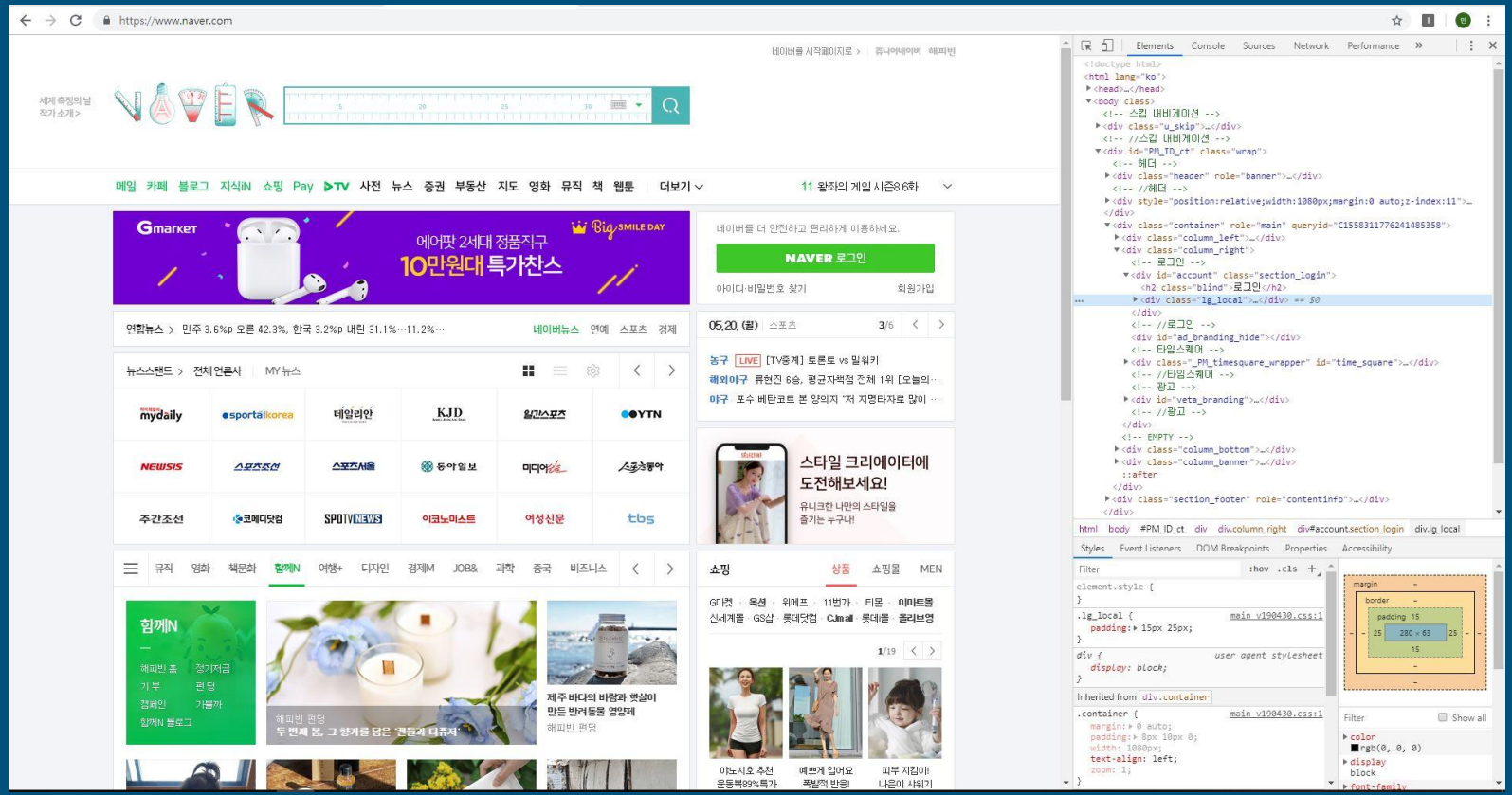
비전공자를 위한 파이썬 기초

한림대학교 영어영문학과 정민

웹 크롤링

- 웹에 기록된 데이터를 긁어 오는 것.
- BeautifulSoup, Selenium이 많이 사용됨.
- 타인의 데이터를 가져오는 것이기 때문에 저작권에 유의해야함.
- 웹크롤링을 하기 위해 해야할 것.
  - 크롤링할 데이터의 html 코드를 분석
  - 데이터들의 유사성을 찾고 반복하여 크롤링할 방법 찾기
  - 적절한 모듈을 사용하여 사용.

크롬 개발자 도구


https://www.naver.com
Elements
Console
Sources
Network
Performance
NAVER
메일 카페 블로그 지식iN 쇼핑 Pay TV 사전 뉴스 증권 부동산 지도 영화 뮤직 책 웹툰 더보기
Gmarket
에어팟 2세대 정품직구
10만원대 특가찬스
NAVER 로그인
스타일 크리에이터에 도전해보세요!
함께N

크롬 개발자 도구

- F12 혹은 ctrl+shift+i를 눌러 개발자 도구를 통해 html코드를 확인 할 수 있음.
- 특정 데이터의 코드를 확인하고 싶을 때는 ctrl+shift+c를 사용하면 됨.

Selenium

- 웹 드라이버를 제어하는 모듈.
- 웹 크롤링에 많이 사용됨.

from selenium import webdriver

실제 코드 예제

```
from selenium import webdriver
import csv
import time

# [create csv file, open]
f = open('cio.csv', 'w', encoding='utf-8_sig',  newline='')
csv_writer = csv.writer(f)

# [Chrome driver set]
path="C:\\Users\\BowlMin\\PycharmProjects\\driver\\chromedriver"
driver = webdriver.Chrome(path)

# [brower_on]
driver.get("http://www.ciokorea.com/t/2996/%EB%B9%85%20%EB%8D%B0%EC%9D%B4%ED%84%B0")



driver.implicitly_wait(3)

# [contents save]
for q in range(5):
    time.sleep(3.0)
    for w in range(1,16):
        try:
            title =
driver.find_element_by_xpath('/html/body/div[3]/div[2]/div/div[1]/div[1]/div['+str(w)+']/div[1]/h4/a')
            hash =
driver.find_element_by_xpath('/html/body/div[3]/div[2]/div/div[1]/div[1]/div['+str(w)+']/div[3]/div[1]/
div[1]')
            date =
driver.find_element_by_xpath('/html/body/div[3]/div[2]/div/div[1]/div[1]/div['+str(w)+']/div[3]/div[1]/
div[2]')
            contests =
driver.find_element_by_xpath('/html/body/div[3]/div[2]/div/div[1]/div[1]/div['+str(w)+']/div[3]/div[2]'
)

            title_text = title.text
            hash_text = hash.text
            date_text = date.text
            contests_text = contests.text

            csv_writer.writerow([title_text, hash_text, date_text, contests_text])
        except:
            pass


 driver.find_element_by_xpath('/html/body/div[3]/div[2]/div/div[1]/div[2]/div[1]/ul/li['+str(q+2)+']/a'
).click()
f.close()
```