



SW멘토링 프로그램

비전공자를 위한 파이썬 기초
한림대학교 영어영문학과 정민



웹 크롤링

- 웹에 기록된 데이터를 긁어 오는 것.
- **BeautifulSoup, Selenium**이 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

The image shows a screenshot of the Naver homepage (naver.com) with the Chrome DevTools developer tool open on the right side. The browser's address bar shows the URL https://www.naver.com. The page content includes a search bar, navigation links, a Gmarket banner for AirPods, a news section with various news sources, and a shopping section. The DevTools interface shows the 'Elements' panel with the HTML structure of the page, and the 'Styles' panel with the CSS rules for the selected element, including a visual box model diagram.

Elements Panel (HTML Structure):

```
<!doctype html>
<html lang="ko">
<head></head>
<body class="">
  <!-- 스크립 내비게이션 -->
  <div class="u_skip"></div>
  <!-- //스크립 내비게이션 -->
  <div id="PM_ID_ct" class="wrap">
    <!-- 헤더 -->
    <div class="header" role="banner"></div>
    <!-- //헤더 -->
    <div style="position:relative;width:1000px;margin:0 auto;z-index:11">
      <div class="container" role="main" queryid="C1558311776241485358">
        <div class="column_left"></div>
        <div class="column_right">
          <!-- 로그인 -->
          <div id="account" class="section_login">
            <h2 class="blind">로그인</h2>
            <div class="lg_local"></div>
          </div>
          <!-- //로그인 -->
          <div id="ag_branding_hide"></div>
          <!-- 타임스퀘어 -->
          <div class="PM_timesquare_wrapper" id="time_square"></div>
          <!-- //타임스퀘어 -->
          <!-- 광고 -->
          <div id="veta_branding"></div>
          <!-- //광고 -->
          <!-- EMPTY -->
          <div class="column_bottom"></div>
          <div class="column_banner"></div>
        </div>
        <div class="section_footer" role="contentinfo"></div>
      </div>
    </div>
  </body>
</html>
```

Styles Panel (CSS Rules):

```
Filter: #PM_ID_ct div div.column_right div#account.section_login div.lg_local
element.style {
}
.lg_local {
  padding: 15px 25px;
}
div {
  display: block;
}
Inherited from div.container:
.container {
  margin: 0 auto;
  padding: 5px 10px 0;
  width: 1000px;
  text-align: left;
  zoom: 1;
}
```

Visual Box Model Diagram:

- margin: 0
- border: 1px solid black
- padding: 15px
- width: 280px
- height: 63px
- border: 1px solid black

크롬 개발자 도구

- F12 혹은 `ctrl+shift+i`를 눌러 개발자 도구를 통해 html코드를 확인 할 수 있음.
- 특정 데이터의 코드를 확인하고 싶을 때는 `ctrl+shift+c`를 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()
```