새로운 기법: XPATH를 이용해서 크롤링하기

- XPath: XML 문서의 특정 요소나 속성에 접근하기 위한 경로를 지정하는 언어

핵심 정리

1. 태그와 속성으로 선택하기

```
crawling_data = soup.find('h1')
crawling_data = soup.find(id='title')
crawling_data = soup.find('p', class_='cssstyle')
crawling_data = soup.find('p', attrs = {'align': 'center'})
```

핵심 정리

2. CSS Selector로 선택하기

```
crawling_data = soup.select('html > title')
crawling_data = soup.select('div.article_view')
crawling_data = soup.select('#harmonyContainer')
crawling_data = soup.select('div#mArticle  div#harmonyContainer')
```

그리고 XPath: 참고!

3. XPath로 선택하기

> BeautifulSoup 에서는 지원하지 않음
> Selenium 과 PhantomJS에서만 사용

새로운 기법: XPATH를 이용해서 크롤링하기

- / : 절대경로를 나타냄
- // : 문서내에서 검색
- //@href : href 속성이 있는 모든 태그 선택
- //a[@href='http://google.com'] : a 태그의 href 속성에 http://google.com 속성값을 가진 모든 태그 선택
- (//a)[3] : 문서의 세 번째 링크 선택
- (//table)[last()] : 문서의 마지막 테이블 선택
- (//a)[position() < 3] : 문서의 처음 두 링크 선택
- //table/tr/* 모든 테이블에서 모든 자식 tr 태그 선택
- //div[@*] 속성이 하나라도 있는 div 태그 선택

> XPATH 문법 상세 참고

```python
from selenium import webdriver

# driver = webdriver.PhantomJS('C:/dev_python/phantomjs-2.1.1-windows/bin/phanto
driver = webdriver.PhantomJS('/usr/local/Cellar/phantomjs/2.1.1/bin/phantomjs') #
driver.get('http://v.media.daum.net/v/20170922175202762')

title = driver.find_element_by_xpath("//title") # 문서내의 어떤 태그든지 가능

# head 태그 안에 있는 title 정보는 get_attribute('text') 메서드로 추출할 수 있습니다.
print (title.get_attribute('text'))
driver.quit()
```

```
# 문서 내 태그 검색
title = driver.find_element_by_xpath("//title") # 문서내의 어떤 태그든지 가능

# 절대경로
title = driver.find_element_by_xpath("/html/head/title")

# html 태그 내에서 다시 검색
title = driver.find_element_by_xpath("/html//title")

# soup.find('h3', attrs = {'class' : 'tit_s'})

title_content = driver.find_element_by_xpath("//h3[@class='tit_view']")
```

실전 예제1

- 페이스북 로그인해보기 (XPATH 사용해보기)

```
from selenium import webdriver
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
from selenium.webdriver.common.by import By
import time

driver = webdriver.Chrome('C:/dev_python/Webdriver/chromedriver.exe') # webdrive
driver.get("https://www.facebook.com/") # facebook 이동

user_name = [-------]
password = [-------]
email_id = [-------]
password_id = [-------]
login_button = [-------]
```

실전 예제1

- 페이스북 로그인해보기 (XPATH 사용해보기)

```
email_tag = WebDriverWait(driver, 10).until(EC.presence_of_element_located((By.I
password_tag = WebDriverWait(driver, 10).until(EC.presence_of_element_located((B
login_button_tag = WebDriverWait(driver, 10).until(EC.presence_of_element_locate



email_tag.clear()
email_tag.send_keys(user_name)
password_tag.clear()
password_tag.send_keys(password)
login_button_tag.click()
driver.quit()
```

실전 예제2

- https://www.seeko.co.kr/zboard4/zboard.php?id=mainnews 에서 제목과 조회수를 최신 순 10개를 가져와서 기사제목과 조회수 출력하기

```
from selenium import webdriver

driver = webdriver.Chrome('C:/dev_python/Webdriver/chromedriver.exe')
driver.get("https://www.seeko.co.kr/zboard4/zboard.php?id=mainnews")

article_data = list()
titles = driver.find_element_by_xpath([-------])
hits = driver.find_element_by_xpath([-------])

for num in range(10):
    article_data.append([titles[num].text, hits[num].text])

print (article_data)
```