학습 정리

팀	이진조	구성원	이규진, 조현진

일정	발제자	주제
05/29	조현진	섹션2. 파이썬 기초 스크랩핑

주요 내용 요약

섹션 2. 파이썬 기초 스크랩핑

1. 스크랩핑 전 크롬 개발자 도구에서 알아야 할 것들!

- 크롬 개발자 도구 (F12)

1. DOM 구조 분석(요소검사)

2. 선택자(Selector) 추출

1) 개발자 도구 원하는 글을 선택

2) Copy -> Copy selector

3. Console 도구

```
a=[1,2,3,4,5,]
(5) [1, 2, 3, 4, 5]

a.map(
      (i,x) => {
            console.log(i,x);
   }
)
1 0
2 1
3 2
4 3
5 4
```

4. Source - 로딩 한 리소스 분석 및 디버깅

5. 네트워크 탭 및 기타

1) 네트워크 : 페이지에 보이는 모든 리소스를 확인할 수 있다. (F5)

* Preserve log

: 새로고침을 하여도 앞에있었던 파일이 모두 누적된다.

(사이트가 아닌 툴로 진행할때 로그인을 할때 어떤방식으로 넘기는지 확인하기 위해 꼭 필요!!)

2) Memory

3) Performance : 로딩되는 순서와 타이밍을 알수있다.

4) Application : 쿠기값을 확인 및 삭제 가능.

2. 파이썬 urlib을 활용한 웹에서 필요한 데이터 추출하기(1)

* HTML
* 필요한 텍스트, 정보 파싱
* DB, TXT, 엑셀, JSON -> SERVER

* atom에서 한글 쓰기

```
import sys
import io

sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')
sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')

print('hi')
print('한글')
```

* Url을 이용하여 이미지 다운받기(urlretrieve)

```
import sys
import io
import urllib.request # as dw

sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')
sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')

imgUrl
="http://post.phinf.naver.net/MjAxODA4MDFfMjMw/MDAxNTMzMDg4NDAwMTE0.gDPRG
```

ifP9tYmNRSxOvNhKQfi1qsyR4luus9bgZdI6uIg.yzhhIvD7AWlpOb4OK1vOA5F4HLVxCEf
Gb57k9gndK94g.JPEG/Iq_cU6Sac798YMzN22yJSvrEU2GM.jpg" # 다운로드 url
  savePath ="c:/test1.jpg" # 다운 경로

  urllib.request.urlretrieve(imgUrl, savePath)
  # dw.urlretrieve(imgUrl, savePath) 가능

  print("다운로드 완료!")

* HTML 다운(urlretrieve)

  ```python
  import sys
  import io
  import urllib.request

  sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')
  sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')

  htmlURL ="https://google.com"

  savePath2 ="c:/index.html"

  urllib.request.urlretrieve(htmlURL, savePath2)

  print("다운로드 완료!")


* url open

  import sys
  import io
  import urllib.request as dw

  sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')
  sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')

  imgUrl
="http://post.phinf.naver.net/MjAxODA4MDFfMjMw/MDAxNTMzMDg4NDAwMTE0.gDPRG
ifP9tYmNRSxOvNhKQfi1qsyR4luus9bgZdI6uIg.yzhhIvD7AWlpOb4OK1vOA5F4HLVxCEf
Gb57k9gndK94g.JPEG/Iq_cU6Sac798YMzN22yJSvrEU2GM.jpg"
  htmlURL ="https://google.com"

  savePath1 ="c:/test1.jpg"
  savePath2 ="c:/index.html"

  f = dw.urlopen(imgUrl).read()
  f2 = dw.urlopen(htmlURL).read()
  ```

```
saveFile1 = open(savePath1) # w : write, r : read , a : add
saveFile1.write(f)
saveFile1.close()

with open(savePath2, 'wb') as saveFile2:
    saveFile2.write(f2)
# with가 끝나는 부분에서 자동으로 close가 된다.

print("다운로드 완료!")
```

* urlretrieve VS urlopen

urlretrieve	urlopen
*저장 -> open('r') -> 변수에 할당 -> 파싱 -> 저장	* 변수 할당 -> 파싱 -> 저장

2. 파이썬 urlib을 활용한 웹에서 필요한 데이터 추출하기(2)

* Urlopen 파라미터(Parameter) 전달 방법

* Type (자료형 알아보기)

```
print(type({})) # <class 'dict'>
print(type([])) # <class 'list'>
print(type(())) # <class 'tuple'>
```

* decode, geturl, status, getheaders, info, urlparse

* 웹페이지에서 필요한 자료 가져오기

```
import sys
import io
import urllib.request as req

sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')
sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')

url = "http://www.encar.com"

mem = req.urlopen(url)

print("geturl", mem.geturl())
# geturl http://www.encar.com/index.do
print("status", mem.status) # 200(정상), 404(페이지없음), 403(접속안됨), 500()
# status 200
```

```
  print("headers", mem.getheaders())
  # headers [('Date', 'Wed, 29 May 2019 06:33:38 GMT'), ...]
  print("info", mem.info())
  # info Date: Wed, 29 May 2019 06:34:07 GMT
  # ...
  print("code", mem.getcode())
  # code 200
  print("read", mem.read())
  # 페이지의 모든것을 다가져온다. ()안에 숫자를 넣어 원하는 만큼만 가져올 수 있음
  print("read", mem.read(50).decode("utf-8")) # euc-kr ...

  from urllib.parse import urlparse # 따로 import를 시켜줘야함
  print(urlparse("http://www.encar.com?test=test"))
  # ParseResult(scheme='http', netloc='www.encar.com', path='', params='',
query='test=test', fragment='') 따로


* API를 이용한 추출

  API = "https://api6.ipify.org"

  values = {
      'format': 'json'
  }
  print('before',values)
  # before {'format': 'json'}
  params = urlencode(values)
  print('after',params)
  # after format=json

  url = API + "?" + params
  print("dycjd url", url)
  # dycjd url https://api6.ipify.org?format=json -> 요청한 url과 일치

  reqData = req.urlopen(url).read().decode('utf-8')
  print("출력",reqData)
  # 출력 {"ip":"14.46.141.21"} -> IP 출력

* 행정 자치부 홈페이지를 통한 실습

  import sys
  import io
  import urllib.request as req
  from urllib.parse import urlencode

  sys.stdout = io.TextIOWrapper(sys.stdout.detach(), encoding = 'utf-8')
  sys.stderr = io.TextIOWrapper(sys.stderr.detach(), encoding = 'utf-8')

  API = "https://www.mois.go.kr/gpms/view/jsp/rss/rss.jsp"
```

```
values = {
    'ctxCd': '1001'
}
print('before',values)
params = urlencode(values)
print('after',params)

url = API + "?" + params
print("dycjd url", url)

reqData = req.urlopen(url).read().decode('utf-8')
print("출력",reqData)

# 게시판의 html 코드가 출력된다.
```