학습 정리

팀	빅나물	구성원	이지현, 조영현

일정	발제자	주제
5일차(5/31)	조영현	파이썬 기초 스크랩핑

주요 내용 요약

 # BeautifulSoup 사용법 및 간단 웹 파싱 기초

- 웹 파싱: 궁극적으로 어떤 html파일(태그, 요소, 속성 등으로 구성) 잘 분석해서 내가 원하는 데이터의 위치를 찾는 것.

ex) 이미지 경로, 동영상 경로, 문자, 숫자 등

- BeautifulSoup이용하면 손쉽게 파싱 가능.

- anaconda prompt에서 beautifulSoup 설치

```
pip install beautifulsoup4
conda list #설치완료 확인
```

- urljoin

```
from urllib.parse import urljoin

baseUrl="http://test.com/html/a.html"
print(">>",urljoin(baseUrl,"b.html"))
# >> http://test.com/html/b.html 치환되어서 나옴.
print(">>",urljoin(baseUrl,"sub/b.html"))
# >> http://test.com/html/sub/b.html
print(">>",urljoin(baseUrl,"../index.html"))
# >> http://test.com/index.html 상위로 가서 치환되어서 나옴.
print(">>",urljoin(baseUrl,"../img/img.jpg"))
# >> http://test.com/img/img.jpg
print(">>",urljoin(baseUrl,"../css/img.css"))
# >> http://test.com/css/img.css
```

- beautifulsoup 기초

- prettify() : html 자동 들여쓰기해서 출력.

- soup.html.body.h1: soup 객체 html태그 안의 body 태그 안의 h1태그 선택

- h1.string: 문자열만 출력

- p 태그는 여러 개 있으므로 soup.html.body.p의 경우 첫번째 노드만 가져옴.

  따라서, next_sibling을 이용해 다음 p태그에 접근.

  `"""" """"`의 경우, 각 줄 마지막에 \n이 포함되어 있기 때문에 next_sibling을 하면 \n가 선택됨.

  따라서, next_sibling.next_sibling을 해주어야 다음 태그가 출력됨. 물론 한 줄 쓰기 경우엔 그럴 필요 없음.

- previous_sibling의 경우도 next_sibling과 마찬가지이나 이전 태그로 이동.

- 태그 선택자 이용해 한번에 가져오기

- .find_all("a"): 모든 a태그 가져옴.
- .find_all("a",string="daum"): a 태그 중 string이 조건에 맞는 것만 가져옴.

- .find("a"): a태그 상위 1개만 가져옴.

- .find_all("a",limit=3): "a" 태그 중 상위 3개만 가져옴.

- .find_all(string=["naver","google"]): string에 해당하는 string만 가져옴. 보통 쓸일 x.

- type이 resultset의 경우 for문 이용해서 출력해야함.

- .attrs[key 값]: key 속성의 value 출력

  href=""에서 href가 key ""가 value

- css 선택자 이용해서 조건에 맞는 element 가져오기

- .select("div#main > h1"): div태그에서 id가 main인 것의 하위의 h1

- print('h1',h1.string)
  에러남. 왜냐하면 h1의 type이 list이기 때문에 바로 리스트의 속성에 접근 불가능.

  따라서 하나라도 반복문을 돌려야 함.

- 하지만 번거로움 피하기 위해 .select_one 사용하면 됨. 하나인 경우!

참고 사이트
인프런
파이썬 입문 및 웹 크롤링을 활용한 다양한 자동화 어플리케이션 제작하기
https://www.inflearn.com/course/python-%ED%8C%8C%EC%9D%B4%EC%8D%AC-%E

C%9B%B9-%EB%8D%B0%EC%9D%B4%ED%84%B0-%ED%81%AC%EB%A1%A4%EB%A7%81/dashboard