2021.03.17

[웹 데이터 수집–I]

1. Beautiful Soup 과 parser라이브러리를 이용한 웹 데이터 수집

(1) 필요한 웹문서 가져오기

웹문서를 불러오기 위한 import	**import** requests **from** request_file **import** FileAdapter
수집할 웹 문서 가져오기	s = **requests.Session( )** s.**mount**('file://, **FileAdapter**) res = s.**get**(file:///파일경로/...웹문서명")
Beautiful Soup 모듈 import	**from** bs4 **import** BeautifulSoup
parser 라이브러리로 html태그와 내용 불러오기	soup = **BeautifulSoup**(res.**content**, **'html.parser'**)

(2) Selector API : BeautifulSoup은 가장 일반적으로 사용되는 CSS선택자들을 지원한다.

ex) soup.**select_one**("h1") *# 해당 문서의 h1태그 내용들 중 첫 h1태그만 불러온다.*

soup.**select**("h1") *# 해당 문서의 h1태그 내용들 전부 불러온다. (보통 for문을 돌려서 출력)*

2. requests를 이용한 웹데이터 수집

(1) 내용 수집할 웹사이트 가져오기

anaconda prompt창에서 설치하기	**pip install requests**
설치한 모듈 import	**import** requests
콘텐츠 수집할 사이트 get 요청	**requests.get**("사이트 주소")

(2) response 객체: 요청 결과를 저장한다.

1) response.**status_code**: 상태코드 출력
2) response.**content**: 응답 내용을 바이트 단위로 출력
3) response.**text**: 응답 내용의 텍스트만 출력
4) response.**encoding= ' '** : 인코딩 방식 지정
5) response.**json( )**: 응답 내용을 json 형식으로 출력

*** 상태 코드(숫자로 표현) ***

- 100번 영역: 정보 전송
- 200 (성공) / 201(post요청) / 204 (전송할 데이터 없음)
- 300번 영역: 리다이렉션
- 401 (사용자인증) / 403(접근 권한 없음) 404 (요청한 url 없음)
- 500번 영역: 서버측 오류