24.Python Scraper

1. Python Scraper I

1] Web Scaper개념

1) Web 주소(Domain 이름)을 받고 HTML data를 받아옴

2) data를 Parsing하여 원하는 정보를 획득

3) 원하는 정보를 저장

4) 필요하다면 다른 Page에서도 이 작업 반복 (**Crawling**)

5) Scraping을 수행하는 프로그램을 일반적으로 봇 또는 Robot이라 함

2] 필요모듈 **(3rd Party)**

1) **requests**

① headers : header Tag내용

② text : 해당 site의 source 보기 내용과 동일

2) **BeautifulSoup**

- **정형화되지않은 HTML(XML)을 Parsing하기 좋게 Python객체로 돌려준다**
- 잘못된 HTML Well-Formed 수정 반환
- Web Browser의 개발자 도구를 사용, 특정Tag에 접근하는 방법을 획득
- Coding : from bs4 import BeautifulSoup(python Console 에서 확인)
- 관련 site : https://www.crummy.com/software/BeautifulSoup/bs4/doc/
- NavigableString 객체

① .string : tag를 모두 Return

② .text : 모든 글자를 Return

3) **lxml**

- html.parser(기본) 보다 좀 더 강력한 lxml(설치필요)많이 사용
- 전문가들이 사용 권장

2. Python Scraper 핵심 BeautifulSoup 과 find

1] BeautifulSoup 개념

- HTML(XML)을 Parsing하기 좋게 Python객체로 Return
- HTML Page에서 원하는 Tag를 다양한 속성에 따라 쉽게 Filtering 가능

2] 주요 함수

1) find() : 첫번째 나타나는 tag반환

① find(tag,Attributes) # tag.find('div')

② find(Attributes)

③ find(class='box') # class명으로 찾기

2) find_all() : 전체를 List 형태로 반환

앵크 tag중에서 href 속성이 'link2'인 속성

① find_all (tag,Attributes) # tag.find_all('a' , href='link2') ,

② find_all (Attributes)

3) 예시

- find (re.compile('^b')) # b로 시작하는 첫번째 Tag
- find_all(re.compile('t')) # t를 포함한 모든 Tag
- find_all(href=re.compile('[₩w]{3}']

4) http://www.pythonscraping.com/pages/page3.html

- 웹 스크렙핑 위해 만들어 놓은 page

3] scraping은 전체적으로 합법적인것은 아님

- robots.txt로 보호되고 있는 자료 및 데이터 베이스를 타 검색 로봇이 수집하는 것을 불허
- scraping 하지 말라고 공지한것 scraping 하면 위법 (주의필요)

예시 : https://www.naver.com/robots.txt

3. Python Scraper 핵심 BeautifulSoup 과 select

1] BeautifulSoup 개념

- HTML(XML)을 Parsing하기 좋게 Python객체로 Return
- HTML Page에서 원하는 Tag를 다양한 속성에 따라 쉽게 Filtering 가능

2] 주요 함수

1) select() : css 표기법에 나타나는 tag반환

2) 예시 (CSS Select)

- soup.select ("head > title") # head tag 밑에 title tag 검색 , 결과는 list
- soup.select (".general") # class 명으로 찾기
- soup.select ("a#link1") # anchor tag 이며 id가 link1 찾기
- soup.select ('a[href^="http://exam.com/"]') # anchor tag 이며 href속성이 http://exam.com/ 특정문자열로 시작
- soup.select ('a[href$="title') # anchor tag 이며 href속성이 title 특정문자열로 끝남

4) http://example.webscraping.com/

- 웹 스크렙핑 위해 만들어 놓은 page
- site 모든 내용을 Excel로 저장 (전체 250 국가 전체 정보 가져오기)
  - ① 현재 국가 Click 시 해당국가 상세정보
  - ② next Click 시 다른 국가 계속 나타남(전체 25 Page)