

데이터 수집 방법 : Crawling ▼



통합검색

20' CMD
정은희, 김희동

2020. 08. 07




목차

1. 데이터 수집 방법

2. Crawling 실습



1. 데이터 수집 방법

수집하다² (蒐集하다)  ★

[동사] 취미나 연구를 위하여 여러 가지 물건이나 재료를 **찾아 모으다**

[유의어] 채집하다, 모으다

Parsing

Crawling

Scraping



1. 데이터 수집 방법

Parsing

“일련의 문자열을 의미있는 토큰(Token)으로 분해하고 그것들로 이루어진 파스 트리로 만드는 과정” => 어떤 data를 원하는 form으로 만들어 내는 것!

예) JSON 파싱, XML 파싱(문서를 구성하는 태그를 컴퓨터가 알아 볼 수 있도록 바꿔주는 과정)

Crawling

”웹 페이지의 내부 링크를 따라 콘텐츠를 색인하고 검색하기 위해 인터넷을 체계적으로 검색하는 행위” = 스파이더링

활용) 검색엔진 유지관리

Scraping

“특정 웹 사이트 또는 페이지에서 특정 정보를 검색 및 추출하는 것“

활용)뱅크샐러드 - 고객의 카드사, 은행, 보험정보들을 각 사이트에서 추출하여 통합 조회



1. 데이터 수집 방법 - Crawling

Robots.txt

“로봇 배제 표준(robots exclusion standard), 로봇 배제 프로토콜(robots exclusion protocol)은 웹 사이트에 로봇이 접근하는 것을 방지하기 위한 규약으로, 일반적으로 접근 제한에 대한 설명을 robots.txt에 기술한다.”

- ✓ 권고안으로, 반드시 지킬 필요는 없음
- ✓ Robots.txt 파일은 항상 웹 사이트의 루트 디렉토리에 위치해야 함
- ✓ 확인 방법 : www.example.com/robots.txt



1. 데이터 수집 방법 - Crawling

Robots.txt

특정 디렉토리의 접근을 허가하려면

```
User-agent: 제어할 로봇의 User-Agent  
Allow: /foo/bar/
```

특정 디렉토리의 접근을 차단하려면

```
User-agent: 제어할 로봇의 User-Agent  
Disallow: /foo/bar/
```

모든 문서에 대해 접근을 허가하려면 (사실상 의미는 없다.)

```
User-agent: *  
Allow: /
```

사용 중인 사이트: [디시인사이드](#) 등.

모든 문서에 대해 접근을 차단하려면

```
User-agent: *  
Disallow: /
```

사용 중인 사이트: [쿨엔조이](#) 등.

모든 문서에 대해 접근을 차단하고, 첫 페이지에 대해서만 허가

```
User-agent: *  
Disallow: /  
Allow : /$
```

사용 중인 사이트: [네이버](#) 메인화면^[1] 등.



1. 데이터 수집 방법 - Crawling

”웹 페이지의 내부 링크를 따라 콘텐츠를 색인하고 검색하기 위해 인터넷을 체계적으로 검색하는 행위”



웹 페이지를 자동으로 탐색하고,
탐색한 정보들을 그대로 복사 및 가공하기 위한 기술



1. 데이터 수집 방법 - Crawling

활용 - 검색엔진 유지관리

Google	Googlebot
Google image	Googlebot-image
Msn	MSNBot
Naver	Yeti ^[2]
Daum	Daumoa

Robots.txt 예시 - Googlebot 로봇에 특정 디렉터리 접근을 차단

```
User-agent: Googlebot  
Disallow: /private/
```




2. Crawling 실습

colab

코랩

<https://colab.research.google.com/notebooks/welcome.ipynb>

사용법

https://tykimos.github.io/2019/01/22/colab_getting_started/



2. Crawling 실습

Colab 기본 사용법

1. 구글 로그인
2. 파일 - 새 노트
3. 작성
4. 실행(좌측 재생버튼 / 런타임 - 실행 / Ctrl + Enter)



2. Crawling 실습

```
import requests  
from bs4 import BeautifulSoup
```

- **requests** - HTTP 요청을 보냄
- **BeautifulSoup** - html 코드를 Python이 이해하는 객체 구조로 변환



2. Crawling 실습

```
import requests
from bs4 import BeautifulSoup

req=requests.get("http://www.naver.com")
html = req.text
#html

#BeautifulSoup으로 html소스를 python 객체로 변환
soup = BeautifulSoup(html, 'html.parser')

#select 메서드 : CSS Selector를 이용해 조건과 일치하는 모든 객체들을 List로 반환
links=soup.select('ol > li > a')
```

- ✓ **req.text** - html을 text형태로 반환
- ✓ **BeautifulSoup(html 소스코드, 이용할 paresr)**



2. Crawling 실습



구글 실시간 검색어 순위

실습 - 구글 실시간 검색어 순위 10위까지 출력하기

구글 실시간 검색어 순위 사이트

<http://rank.epizy.com/?i=1>

1. 삼성
2. 갤럭시 버즈 라이브
3. 코로나19 예방
4. 레바논
5. 류호정
6. 디지털교도소
7. 다만 악에서 구하소서
8. 김호중
9. 황정민
10. 손정우
11. 용혜인
12. 이지현
13. V4
14. 김진애
15. 참피디
16. 문복희
17. 임슬옹 사망
18. 쫄쫄
19. 박상철
20. 김재우



2. Crawling 실습

+)

-Parameter를 이용해서 여러 사이트에 접속

-Session 값 / API 키 값을 통해서 로그인 필요한 데이터 추출 ex)에브리타임

감사합니다.

