네트워크란

컴퓨터는 태초에 복잡한 계산등의 업무를 처리함

이것만으로는 사용하기 아까웠기 때문에, 컴퓨터를 연결해보기로 함

일정 지역의 네트워크를 묶어 근거리 지역 네트워크 (Local Area Network, LAN)가 탄생함

이후 범 지구적 네트워크인 인터넷이 탄생한다

그럼 웹이 뭐냐면

이 인터넷에서 정보를 교환할 수 있는 하나의 플랫폼. World Wide Web 인 WWW의 웹이다

인터넷은 망이고, 웹은 플랫폼이다

컴퓨터는 클라이언트와(사용자)와 서버(제공자)가 요청을 주고받으며 업무를 처리한다 (패킷교환)

이것을 우리가 HTTP라고 함

Hypertext transfer protocol

HTTP는 Host,Resource,Method 등을 정보로 하여 전달됨

Head에는 송,수신자의 정보가

Body에는 내용이 들어있음

클라이언트가 서버로 보낼 때

GET (요청사항) /(최상단의 정보) HTTP 1.1

Host: [www.programmers.com](http://www.programmers.com) (받고싶은 정보)

User-Agent: Mozilla/5.0 (사용자 정보)

…(혹시 바꾸고싶거나 하는 정보들)

서버에서 클라로 답장

HTTP / 1.1 200 OK(잘 받았습니당)

…

<html>…</html>(돌려줄 정보, 상황에 따라 다른 정보나 시그널이 올 수도 있음 EX 404)

HTML의 특징들

HTML은 가장 바깥에 <html> 태그로 감싸져있다

HTML은 여러 태그들로 내용을 묶어서 글의 형식을 지정한다

웹 스크래핑

웹 페이지들로부터 우리가 원하는 정보를 추출

웹 크롤링

크롤러를 이용해서 웹 페이지의 정보를 인덱싱

웹 페이지의 링크를 타고 타고 타고 다니며 서치함

정리 :

웹 스크래핑 -> 특정한 목적으로 특정 웹페이지 (데이터 추출)

웹 크롤링 -> URL을 타고다니며 반복적으로 데이터를 가져오는 과정 (데이터 색인)

웹 스크래핑/크롤링으로 어떤 목적을 달성하고자 하는지, 서버에 영향을 미치지는 않는지 고려해야 할 것

그래서 사람들은 웹 크롤링에 규칙이 생김

모든 유저 에이전트에게 접근 거부 방법(요청 보낸사람)

User-agent : \*

Disallow : /

반대로 모든 유저 에이전트에게 접근 허용

User-agent : \*

Allow : /

특정 키워드로 특정 대상에게 허용

User-agent : yiReal

Disallow : /

저사람만 차단

Robots.txt

기본 주소가 있을떄 robots.txt를 붙여주면 참조 가능

Ex) [www.naver.com/robots.txt](http://www.naver.com/robots.txt)

User-agent : \*

Disallow : /

Allow : /$

모든 유저에 대하여 모든 요청 거절

단 [www.naver.com/](http://www.naver.com/) 까지는 허용함

파싱이란

브라우저의 렌더링 엔진은 웹 문서를 로드한 후

HTML을 head와 body부분으로 트리 형태로 구분하여 나누는 것

이를DOM (Document Object Model) 로 부른다

실제로는 엄청나게 복잡한 트리의 구조로 존재한다.

아래의 노드들은 모두 각자의 객체로서 취급하여 문서를 더욱 편리하게 관리 가능하다

DOM을 사용하는 이점은 원하는 요소를 동적으로 변경하거나 쉽게 찾을 수 있다.

중요한 점은, 브라우저의 DOM처럼 파이썬으로 HTML을 분석하는 HTML Parser가 필요하다!