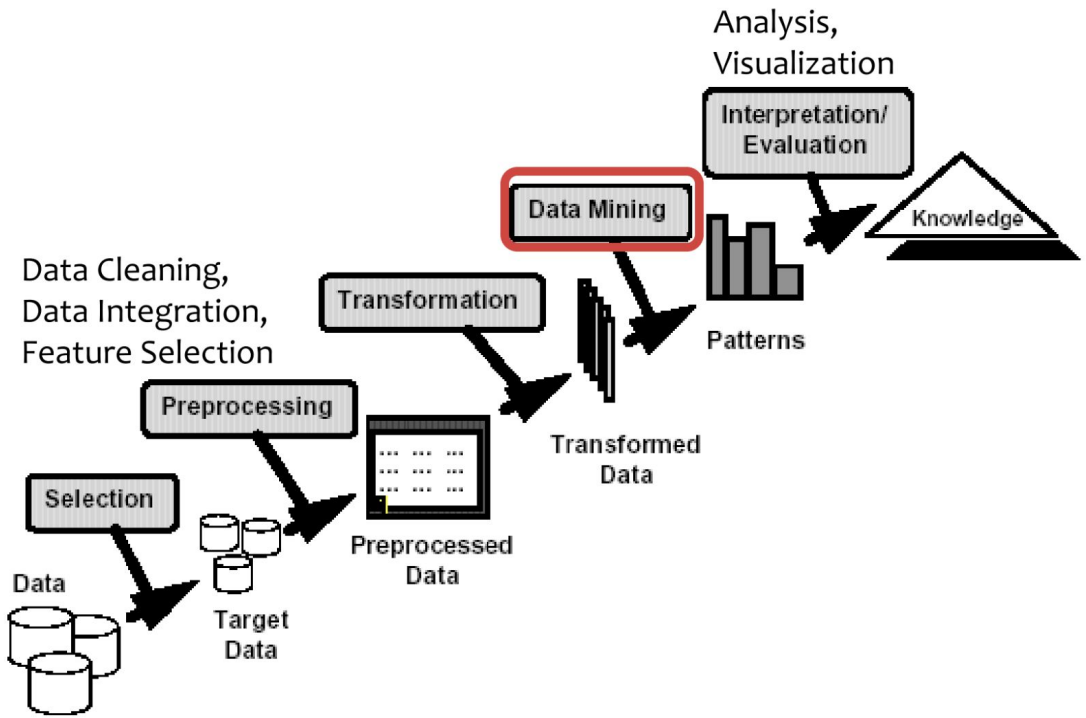
[6/6]
데이터 기초와 크롤링

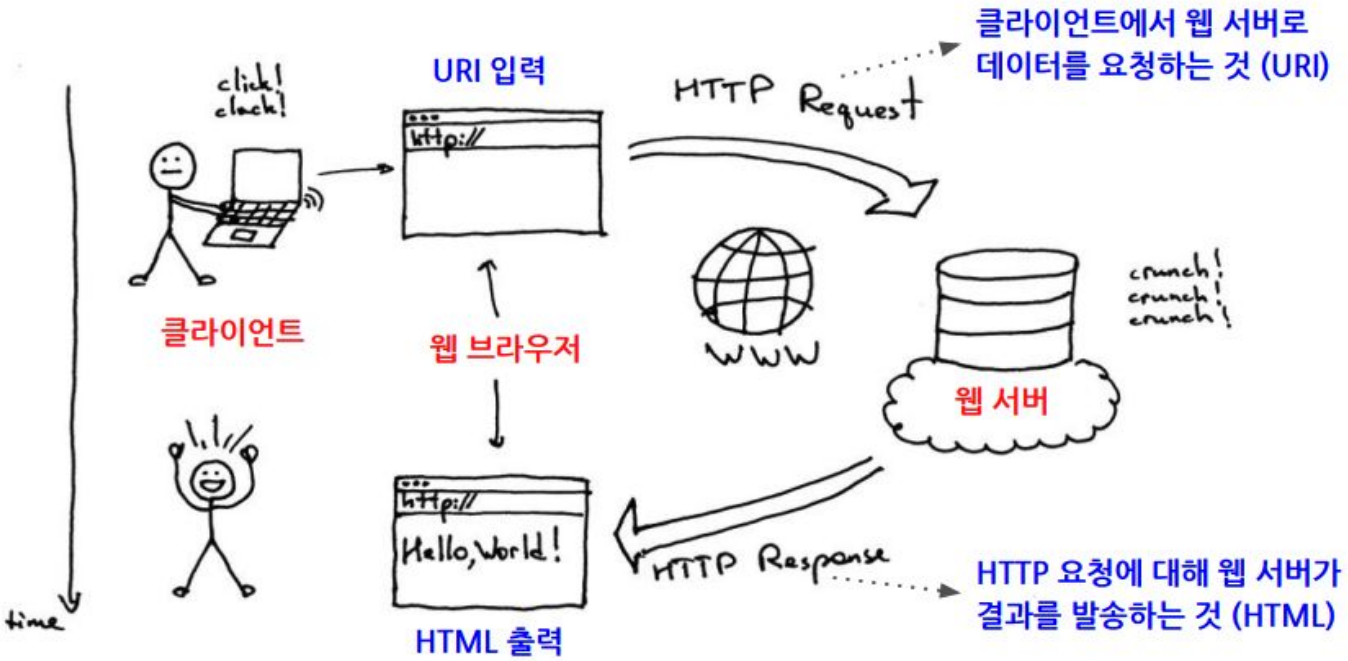
데이터 기초 과목

1. 데이터의 기초
2. 웹 크롤링
3. API를 활용한 공공데이터 수집
4. Pandas 복습
5. 텍스트 데이터의 전처리
6. 데이터 시각화
7. SQL 기초와 데이터 베이스

데이터 수집/모델링 과정


Analysis,
Visualization
Interpretation/
Evaluation
Knowledge
Data Mining
Patterns
Data Cleaning,
Data Integration,
Feature Selection
Transformation
Transformed
Data
Preprocessing
Preprocessed
Data
Selection
Data
Target
Data

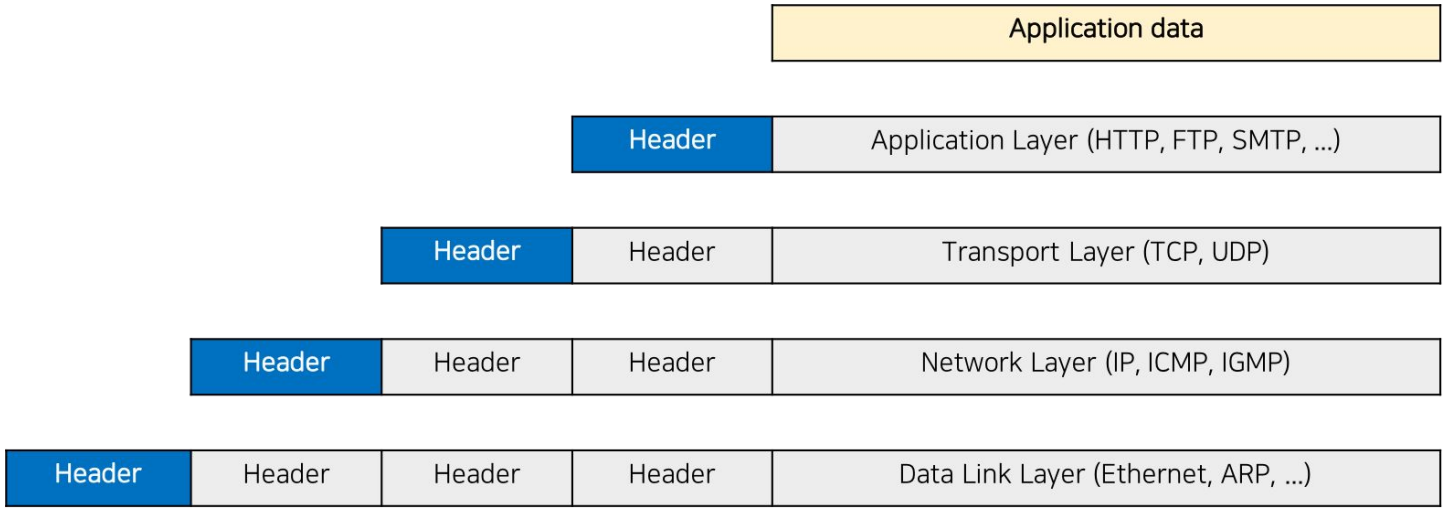
데이터 전달 과정


클라이언트에서 웹 서버로
데이터를 요청하는 것 (URI)
URI 입력
HTTP Request
click!
clack!
http://
클라이언트
웹 브라우저
WWW
crunch!
crunch!
crunch!
웹 서버
time
Hello, World!
HTTP Response
HTTP 요청에 대해 웹 서버가
결과를 발송하는 것 (HTML)
HTML 출력

웹크롤링이란?

- 웹크롤링(Crawling)은 웹페이지에서 보이는 데이터를 필요한 부분만 선택하여 수집하는 행위를 말합니다. 스크레이핑(Scraping)이라고도합니다.
  - 웹크롤링에 사용되는 프로그램을 크롤러라고 합니다.
- IE,크롬과 같은 웹브라우저 상에서 보이는 데이터는 크롤링이 가능하다고 할 수 있습니다. 그러므로 필요로 하는 데이터를 포함하고 있는 웹사이트를 발견하는 것이 웹크롤링의 시작이 됩니다.
- 웹 크롤링 방법은 웹페이지에 따라 서로 다르게 적용해야 합니다. 본 강의를 통해 웹크롤링에 필요한 다양한 방법을 익힐수 있습니다

네트워크 / 웹 구조


Application data
Header
Application Layer (HTTP, FTP, SMTP, ...)
Header
Header
Transport Layer (TCP, UDP)
Header
Header
Header
Network Layer (IP, ICMP, IGMP)
Header
Header
Header
Header
Data Link Layer (Ethernet, ARP, ...)

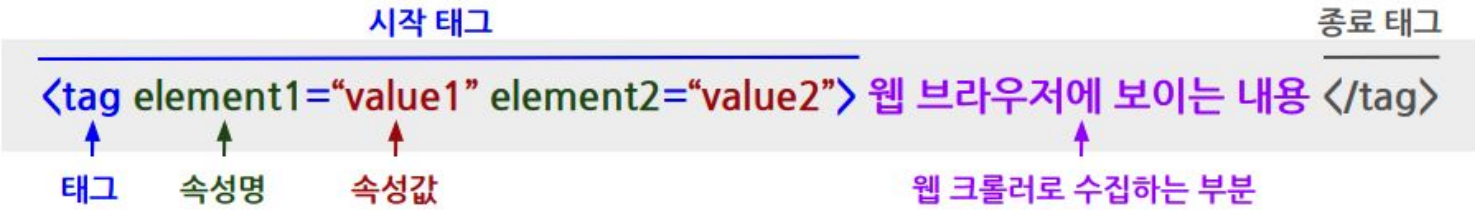
웹 크롤링은 인터넷 검색과 유사

HTML (HyperText Markup Language)

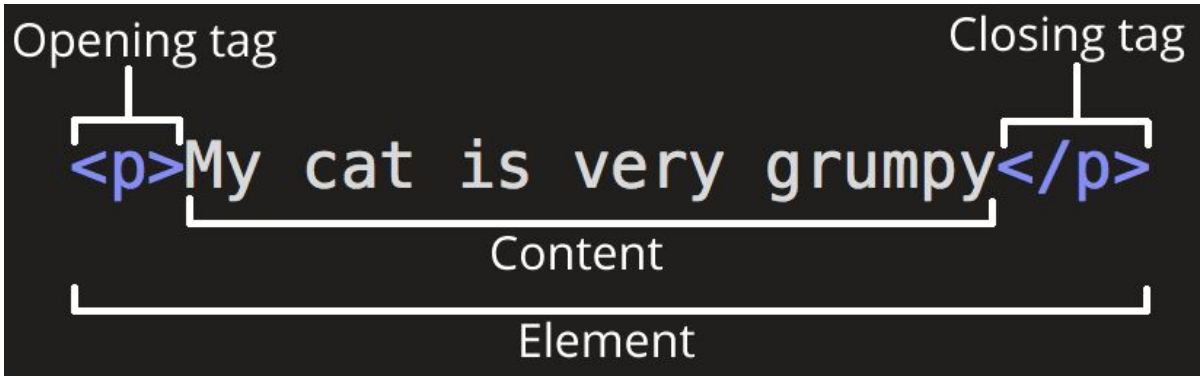
- HTML은 웹 페이지의 제목, 단락, 목록 등 **문서의 구조를 나타내는 마크업 언어**입니다.
- HTML은 꺽쇠 괄호 '< >' 안에 태그로 되어 있는 HTML 요소 형태로 작성됩니다.
- HTML의 디자인을 담당하는 CSS와 웹 브라우저를 제어하는 JavaScript를 함께 사용함으로써 상호작용하는 웹 페이지를 구현할 수 있습니다.

HTML element (요소)

- HTML 요소는 HTML 문서나 웹 페이지를 구성하는 개별 항목을 의미합니다.
- HTML 요소는 시작 태그와 종료 태그로 작성되며, 그 사이에 내용이 포함됩니다.
- 태그는 꺽쇠 괄호로 감쌉니다. 시작태그에 속성명과 속성값이 포함되고, 종료 태그에는 '/'가 추가됩니다.
- 웹 크롤링은 수집하려는 부분을 포함하는 HTML 요소를 찾는 것이 필수입니다.



HTML


Opening tag
Closing tag
My cat is very grumpy
Content
Element

Crawling & BeautifulSoup

2. 크롤링(crawling) 이해 및 기본

2.1. 크롤링(crawling) 이란?

Web상에 존재하는 Contents를 수집하는 작업 (프로그래밍으로 자동화 가능)

1. HTML 페이지를 가져와서, HTML/CSS등을 파싱하고, 필요한 데이터만 추출하는 기법
2. Open API(Rest API)를 제공하는 서비스에 Open API를 호출해서, 받은 데이터 중 필요한 데이터만 추출하는 기법
3. Selenium등 브라우저를 프로그래밍으로 조작해서, 필요한 데이터만 추출하는 기법

2.2. BeautifulSoup 라이브러리를 활용한 초간단 예제

HTML의 태그를 파싱해서 필요한 데이터만 추출하는 함수를 제공하는 라이브러리

- BeautifulSoup 라이브러리 페이지
- 설치 방법 : pip install bs4
- 참고: BeautifulSoup 4 API Guide