빅데이터 수집 개요

1. 데이터 수집 개요
2. 빅데이터 수집 방법

1. 빅데이터 수집 개요

- **데이터 수집 정의**
  - '서비스 활용에 필요한 데이터를 시스템의 내부 혹은 외부에서 주기성을 갖고 필요한 형태로 수집하는 활동'
  - 다양한 유형의 데이터(정형, 반정형, 비정형)를 수집하는 것
  - 데이터 수집 기술은 빅데이터 제공 서비스의 품질을 결정하는 중요한 기술
    - 데이터가 존재해도 그것을 수집할 수 있는 기술이 미비하다면 빅데이터 관련 서비스에 대한 품질을 기대하기 어렵다.

1. 빅데이터 수집 개요

- **수집 데이터 형태에 따른 분류**
  - 정형 데이터
    - 관계형 데이터베이스 시스템의 테이블과 같이 고정된 컬럼에 저장되는 데이터
    - 데이터의 스키마를 지원
    - 예 : RDBMS의 테이블들, 스프레드시트 데이터
  - 반정형 데이터
    - 데이터 내부에 정형데이터의 스키마에 해당되는 메타데이터를 갖지고 있음, 파일 형태로 저장
    - 예: URL 형태로 존재(HTML), 오픈 API 형태로 제공(XML, JSON), 로그 형태로 제공(웹 로그, IOT에서 제공하는 센서 데이터)
  - 비정형 데이터
    - 데이터 세트가 아닌 하나의 데이터가 수집 데이터로 객체화 됨
    - 언어 분석이 가능한 텍스트 데이터, 이미지, 동영상 같은 멀티미디어 데이터
    - 예: 이진 파일 형태(동영상, 이미지), 스크립트 파일 형태(소셜 데이터의 텍스트)

1. 빅데이터 수집 개요

수집데이터의 형태와 데이터 수집과의 관계

수집 난이도

형태	특징	난이도
정형 데이터	내부 시스템인 경우가 대부분이라 수집이 쉽다. 파일 형태의 스프레드시트라도 내부에 형식을 가지고 있어 처리가 쉬운 편이다.	하
반정형 데이터	보통 API 형태로 제공되기 때문에 데이터 처리 기술이 요구 된다.	중
비정형 데이터	텍스트 마이닝 혹은 파일일 경우 파일을 데이터 형태로 파싱해야 하기 때문에 수집 데이터 처리가 어렵다.	상

데이터 처리 아키텍처 구성 난이도

형태	특징	난이도
정형 데이터	CRUD가 일어나는 일반적인 아키텍처 구조로 이루어져 있다.	하
반정형 데이터	데이터의 메타구조를 해석해 정형 데이터 형태로 바꿀 수 있는 아키텍처 구조를 수정해야 한다.	중
비정형 데이터	텍스트나 파일을 파싱해 메타구조를 갖는 데이터의 셋형태로 바꾸고 정형 데이터 형태의 구조로 만들 수 있도록 아키텍처 구조를 수정해야 한다.	상

1. 빅데이터 수집 개요

수집데이터의 형태와 데이터 수집과의 관계

데이터의 잠재적 가치

형태	특징	잠재가치
정형 데이터	내부 데이터의 특성상 현실적 가치의 한계상 활용측면에서 잠재적 가치는 상대적으로 낮다.	보통
반정형 데이터	데이터의 제공자가 선별해 제공하는 데이터로 잠재적 가치는 정형 데이터 보다 높다.	높음
비정형 데이터	수집주체에 의해 데이터에 대한 분석이 선행되었기 때문에 목적론적 데이터 특징이 가장 잘 나타나는 데이터이다. 그렇기 때문에 일단 수집이 가능하면 수집주체에게는 가장 높은 잠재적 가치를 제공한다.	매우높음

1. 빅데이터 수집 개요

수집 위치에 따른 데이터 분류

내부 데이터

- 수집 원천 데이터의 데이터 저장소가 내부 시스템에 있는 데이터
- 데이터 제공자와 상호 협약에 의한 의사소통 가능
- 데이터 수집 주기 및 방법은 데이터 제공자(or 기관)와의 협약을 통해 제공 받음
- 수집 실패한 데이터에 대한 재수집 구현 가능

외부 데이터

- 수집 원천 데이터의 데이터 저장소가 외부 시스템에 있는 데이터
- 데이터 제공자와 협약 된 관계가 아니면 상호 의사소통이 불가능
- 데이터 수집을 위해 수집 주기 및 방법에 관한 분석이 필요
- 외부 데이터의 인터페이스 방법은 수집할 항목을 분석해 수집 시스템을 설계
- 협약이 되지 않은 시스템의 경우 수집 실패 시의 대안을 마련 필요
- 가능한 데이터의 전처리 과정 없이 원본 데이터를 수집 후, 수집 시스템에서 처리를 할 수 있도록 인터페이스를 설계하는 것이 바람직하다.

1. 빅데이터 수집 개요

수집 데이터의 위치와 데이터 수집과의 관계

수집 난이도

위치	특징	난이도
내부	데이터의 저장소가 내부에 있으므로 해당 소스 데이터 담당자와 의사소통이 원활하기 때문에 수집난이도가 외부데이터와 비교해 낮다.	하
외부	외부 소스의 경우 해당 소스 데이터 담당자와 의사소통이 어려워 상대적으로 수집난이도가 높다	상

데이터 처리 아키텍처 난이도

위치	특징	난이도
내부	대부분 정형 데이터이므로 일반적인 CRUD처리 아키텍처와 같은 구성이 가능하다.	하
외부	대부분 비정형, 반정형 데이터 형태로 일반적인 아키텍처 구성에 반정형, 비정형 데이터를 처리할 수 있는 아키텍처를 추가해야 한다.	상

데이터 잠재적 가치

위치	특징	난이도
내부	내부 데이터의 특성과 현실적 가치의 한계상 활용 측면에서 잠재적 가치는 상대적으로 낮다.	보통
외부	데이터의 제공자가 선별해 제공하는 데이터나 수집주체에 대한 분석이 이루어진 후 수집을 하는 데이터이기 때문에 데이터의 목적론적 특징이 가장 잘 나타나는 데이터이다. 그렇기 때문에 내부 데이터와 비교할 경우 상대적으로 잠재적 가치가 높다.	높음

1. 빅데이터 수집 개요

- **수집방법의 종류**
  - HTTP 수집
    - 크롤링(Crawling) : 텍스트 정보를 가져오는 수집 기술
    - Open API 수집 기술 : 웹을 운영하는 운영주체가 정보를 제공하는 수집 기술
  - **로그/센서 수집**
    - **로그 수집 기술 :** 데이터 처리 에이전트의 구별을 통해 로그 수집 기술
    - **센서 데이터 수집 기술 :** 실시간 처리가 주를 이루는 머신정보 수집 기술
  - **DBMS 수집**
    - DBMS 수집은 DB에 직접 연결해 데이터를 수집
  - **FTP 수집**
    - 용량 파일을 수집하기 위하 클라이언트 서버 간 연결 및 파일전송, FTP 보안기능이 제공

1. 빅데이터 수집 개요

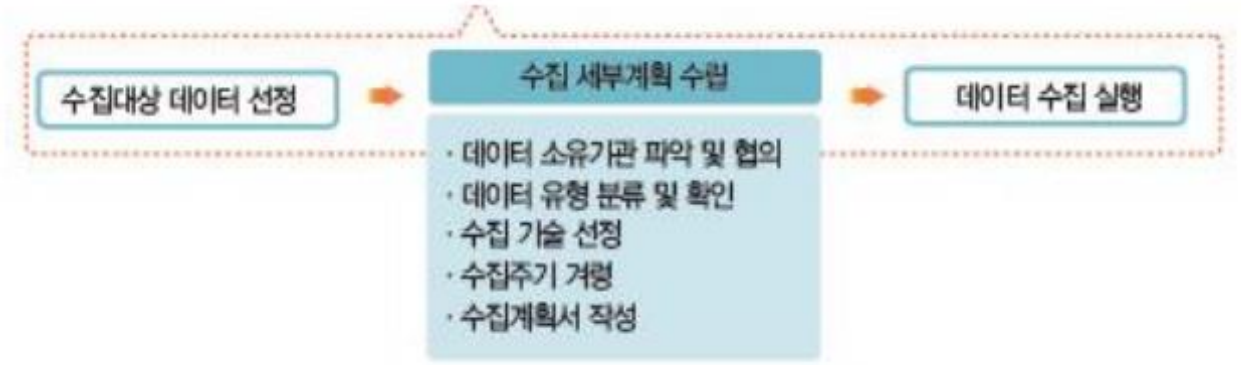
- **수집 대상 데이터 선정**

분석과 서비스 제공할 때 서비스 품질을 결정하는 중요한 단계

수집 가능 ?
사용 가능 ?
이용 목적에 맞는 세부항목 포함 여부
개인정보 침해/ 비용 ?

1. 빅데이터 수집 개요

- 수집 세부 계획 수립



- 데이터 유형 분류

유형	특징	데이터 종류	수집 기술
정형 데이터 (Structured)	– RDBMS의 고정된 필드에 저장 – 데이터 스키마 지원	RDB, 스프레드 시트	ETL, FTP, Open API
반정형 데이터 (Semi-structured)	– 데이터 속성인 메타데이터를 가지며, 일반적으로 스토리지에 저장되는 데이터 파일	HTML, XML, JSON, 웹문서, 웹로그, 센서 데이터	Crawling, RSS, Open API, FTP
	– XML 형태의 데이터로 값과 형식이 다소 일관성이 없음		
비정형 데이터 (Unstructured)	– 언어 분석이 가능한 텍스트 데이터	소셜 데이터, 문서, 이미지, 오디오, 비디오	Crawling, RSS, Open API, Streaming, FTP
	– 형태와 구조가 복잡한 이미지, 동영상 같은 멀티미디어 데이터		

1. 빅데이터 수집 개요

- **데이터 수집 실행**
  - 능동적 데이터 수집
    - 데이터 소유 주체가 수집을 원하는 자에게 능동적 제공
    - 예 : 생산관련 로그, 설문조사
  - VS 수동적 데이터 수집
    - 데이터 소유 주체가 웹페이지에 데이터 공개, 데이터를 원하는 자가 데이터 수집
    - 예 : 웹 로봇, 웹 크롤러

1. 빅데이터 수집 개요

- **데이터 수집 실행**
  - 내부 데이터 수집
    - 내부 파일시스템, DB, 센서
    - ETL(Extraction, Transformation, Loading)
  - 외부 데이터 수집
    - 인터넷으로 연결된 외부에서 데이터 수집
    - Crawling Engine
    - 예: SNS, UCC, 온라인 쇼핑, 검색 등

1. 빅데이터 수집 개요

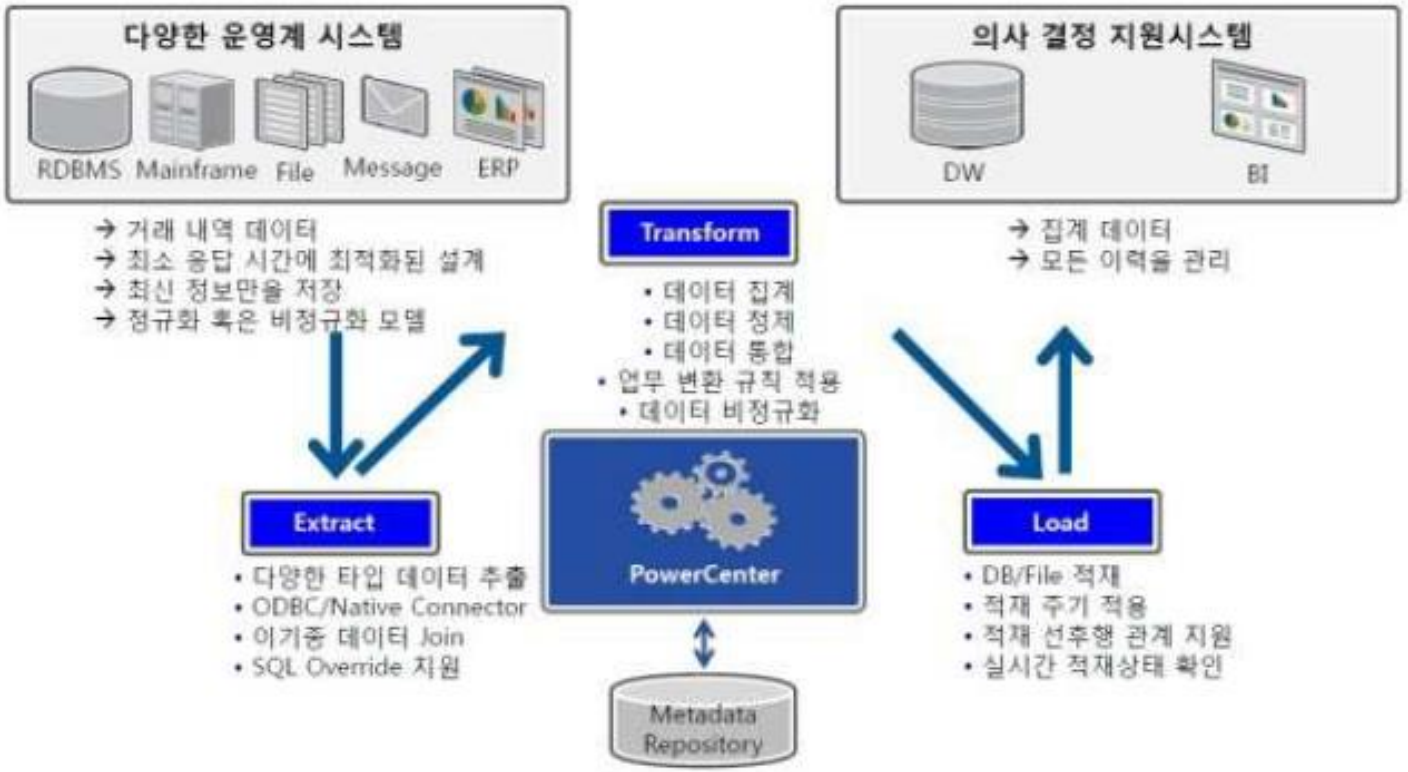
데이터 수집 실행(빅데이터 수집을 위한 변환 및 통합)

- 변환은 정제를 포함
  - 비정형 데이터 정제, 정형데이터에서 측정값이 빠져 있거나, 형식이 다르거나, 내용 자체가 틀린 데이터를 고침

빅데이터 수집을 위한 변환 및 통합	
ETL(Extraction, Transformation, Load)	메인프레임, ERP, CRM, Flat file, Excel 파일 등으로부터 데이터를 추출하여 목표하는 저장소의 데이터 형태로 변형한 후 목표 저장소(DW)에 저장
비정형 → 정형	비정형 데이터는 비구조적 데이터 저장소에 저장하거나 어느 정도 구조적인 형태로 변형하여 저장 ex) Scribe, Flume, chuckwa 등 오픈 소스 솔루션
레거시 데이터와 비정형 데이터간의 통합	데이터를 분석하기 위해서는 수집된 정형의 레거시 데이터와 비정형 데이터간의 통합이 필요 • Sqoop : RDBMS와 HDFS간의 데이터를 연결해 주는 기능으로 SQL 데이터를 Haoop으로 로드하는 도구

1. 빅데이터 수집 개요

- 빅데이터 수집



2. 빅데이터 수집 방법

웹 크롤링과 웹스크랩핑

웹 크롤링	▪ 웹 페이지의 하이퍼링크를 순회하면서 웹 페이지를 다운로드하는 작업
웹 스크래핑	▪ 다운로드한 웹 페이지에서 필요한 콘텐츠를 추출하는 작업 ▪ 웹 페이지를 구성하고 있는 HTML 태그의 콘텐츠나 속성의 값을 읽는 작업

<td>빨강 머리 앤</td>

태그의 콘텐츠

파이썬

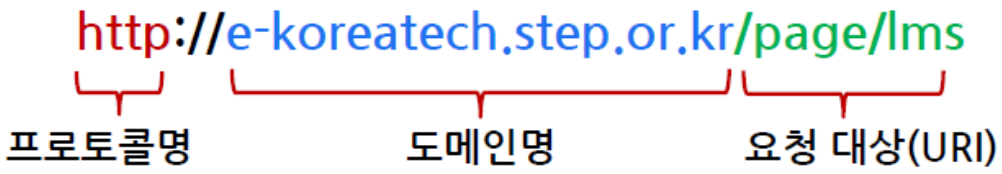
태그의 속성값

2. 빅데이터 수집 방법

웹 크롤링과 웹 스크랩핑

URL(Uniform Resource Locator)

- 네트워크 상에서 자원이 어디 있는지를 알려주기 위한 규약
- 컴퓨터 네트워크와 검색 메커니즘에서의 자원의 위치를 지정하는 문자열



URI(Uniform Resource Identity)

- 웹 사이트에 요청하고자 하는 대상의 패스정보와 파일명으로 구성
- 파일명이 생략되면 디폴트로 index.html 사용

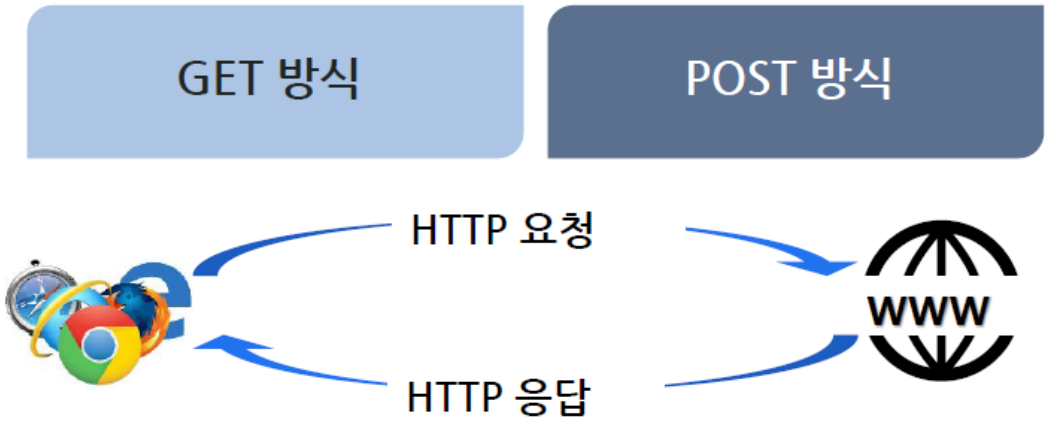
웹 크롤링과 웹 스크래핑

HTTP(HyperText Transfer Protocol)

- **웹상에서 클라이언트와 서버간에 정보를 주고 받을 수 있는 통신규약(프로토콜)**

- URL 문자열을 직접 입력하거나 ,하이퍼링크 텍스트 또는 이미지를 클릭하여 HTML 문서를 주고 받는데 사용
- 디폴트로 80번 포트 사용
- 다른 포트번호를 사용하는 웹 서버에 요청 시 도메인명 뒤에: 기호와 함께 포트 번호 지정
- 웹 클라이언트에서 웹 서버에 HTTP 요청을 전달할 때 요청 방식 명시
- 일반적으로 2 가지 방식(GET, POST) 사용

- **웹 크롤링과 웹 스크래핑**
  - GET 방식과 POST 방식



웹 크롤링과 웹 스크래핑

GET 방식과 POST 방식

GET 방식

- **브라우저에서 직접 요청하려는 페이지의 URL 문자열을 입력하여 요청**
- **하이퍼링크가 설정된 텍스트나 이미지를 클릭하여 요청**

<form> 태그를 통한 요청

- **method 속성 값에 따라서 GET 방식 요청과 POST 방식 요청 모두 가능**

웹 크롤링과 웹 스크래핑

GET 방식과 POST 방식

GET 방식

- Query 문자열 없는 요청과 Query 문자열을 추가한 요청 모두 가능
- Query 문자열이 URL 문자열 뒤에 추가되어 전달

https://movie.daum.net/moviedb/main?movieId=126260

웹 브라우저가 웹 서버에게 요청을 보내면서
함께 전달되는 name과 value로 구성되는 문자열

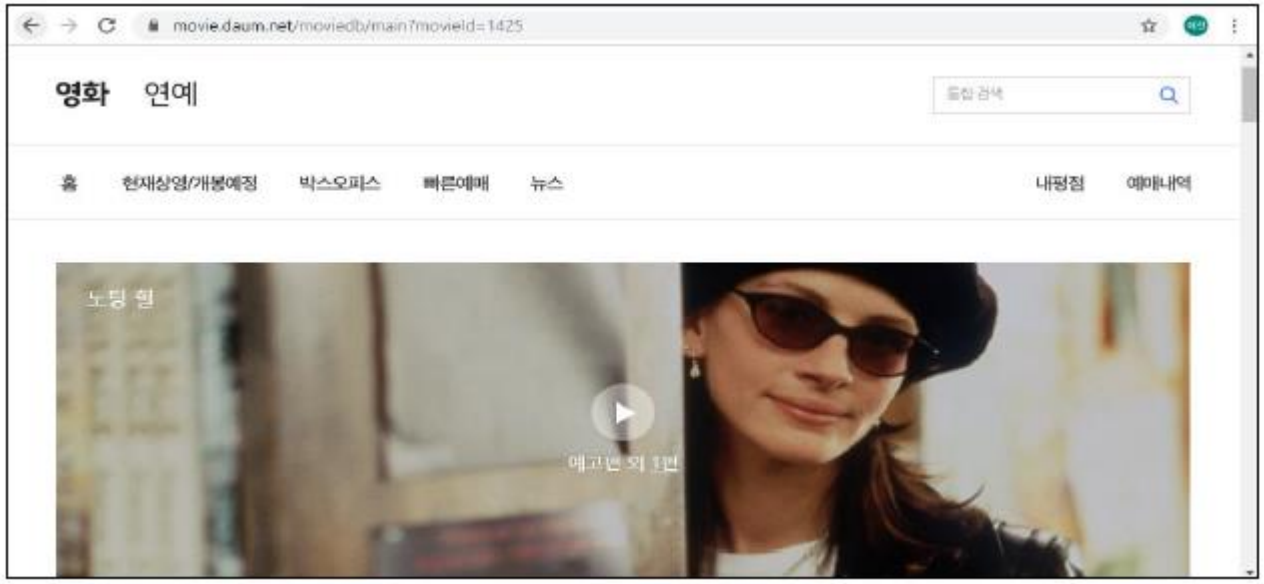
POST 방식

- Query 문자열을 추가한 요청만 가능
- Query 문자열이 요청 바디에 따로 담겨서 전달되므로 요청 URL 문자열에서는 볼 수 없음

웹 크롤링과 웹 스크래핑

크롬(Chrome) 브라우저의 개발자 도구

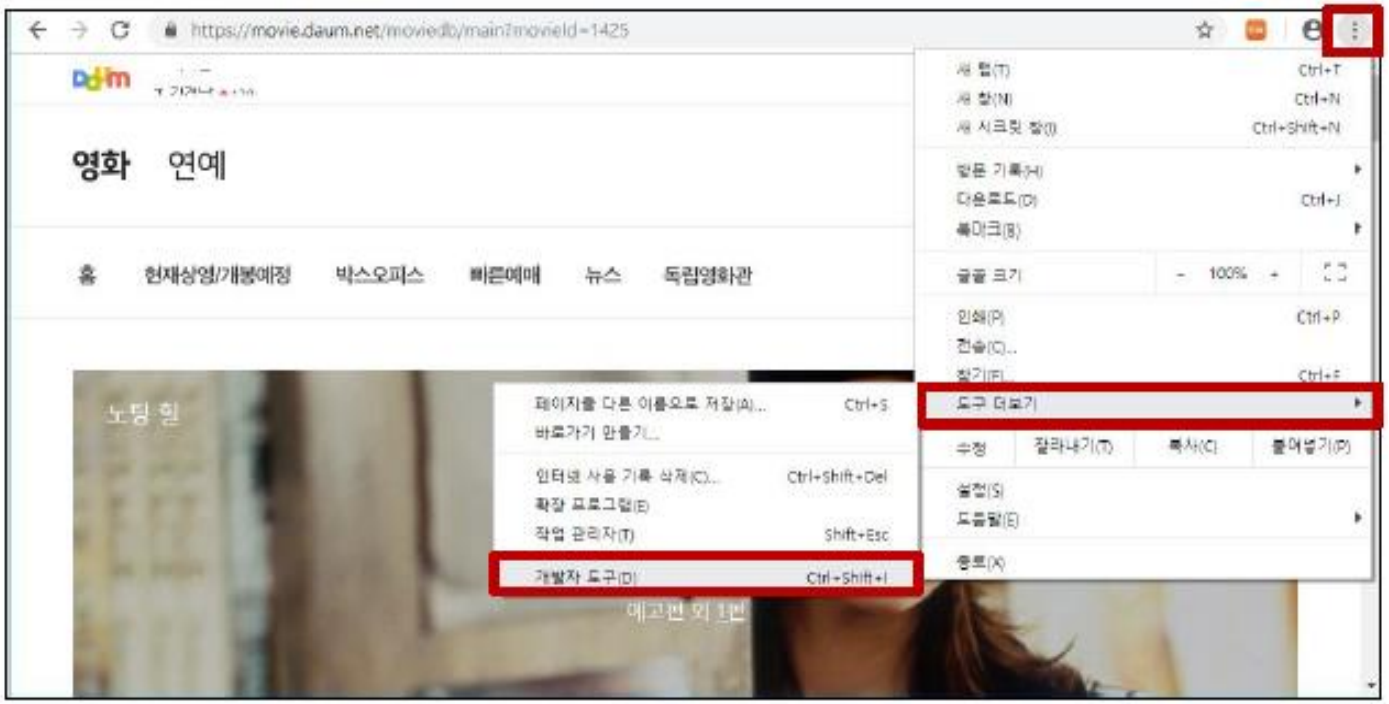
- 크롬 브라우저에서 URL을 입력하고 요청



웹 크롤링과 웹 스크래핑

크롬(Chrome) 브라우저의 개발자 도구

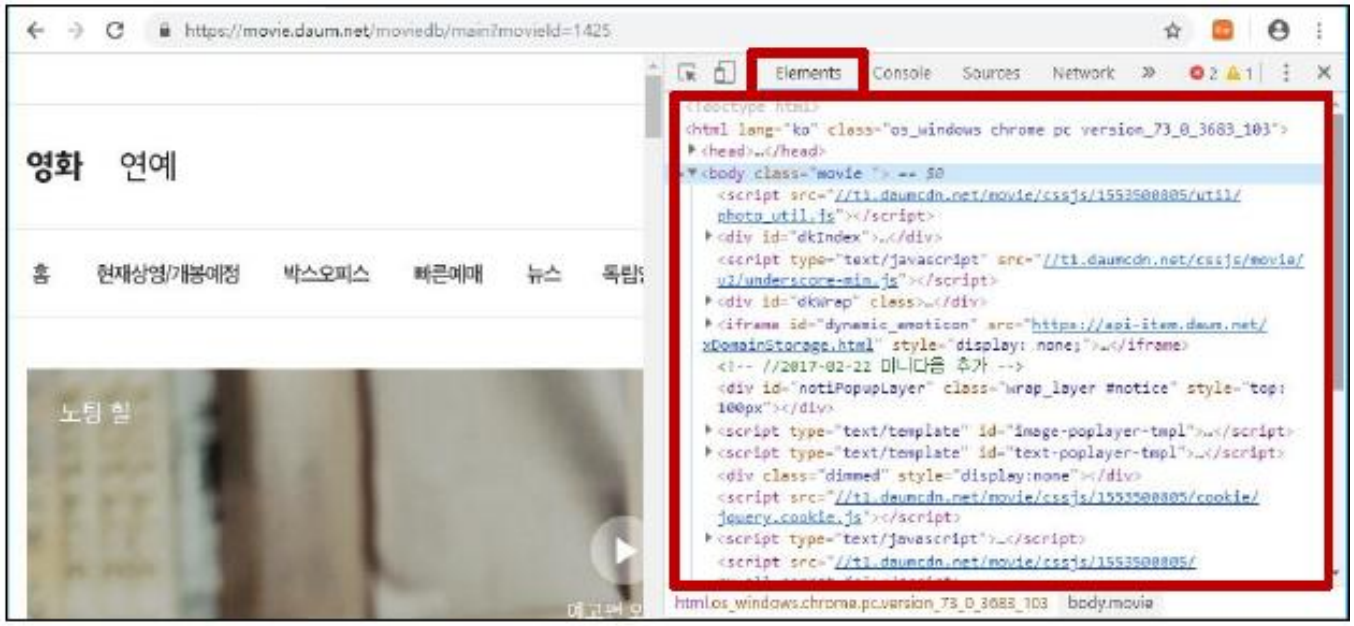
- 오른쪽 상단의 ‘Chrome 맞춤 설정 및 제어’ 메뉴를 클릭한 다음 ‘도구 더보기’ 메뉴의 ‘개발자 도구’ 메뉴 클릭



웹 크롤링과 웹 스크래핑

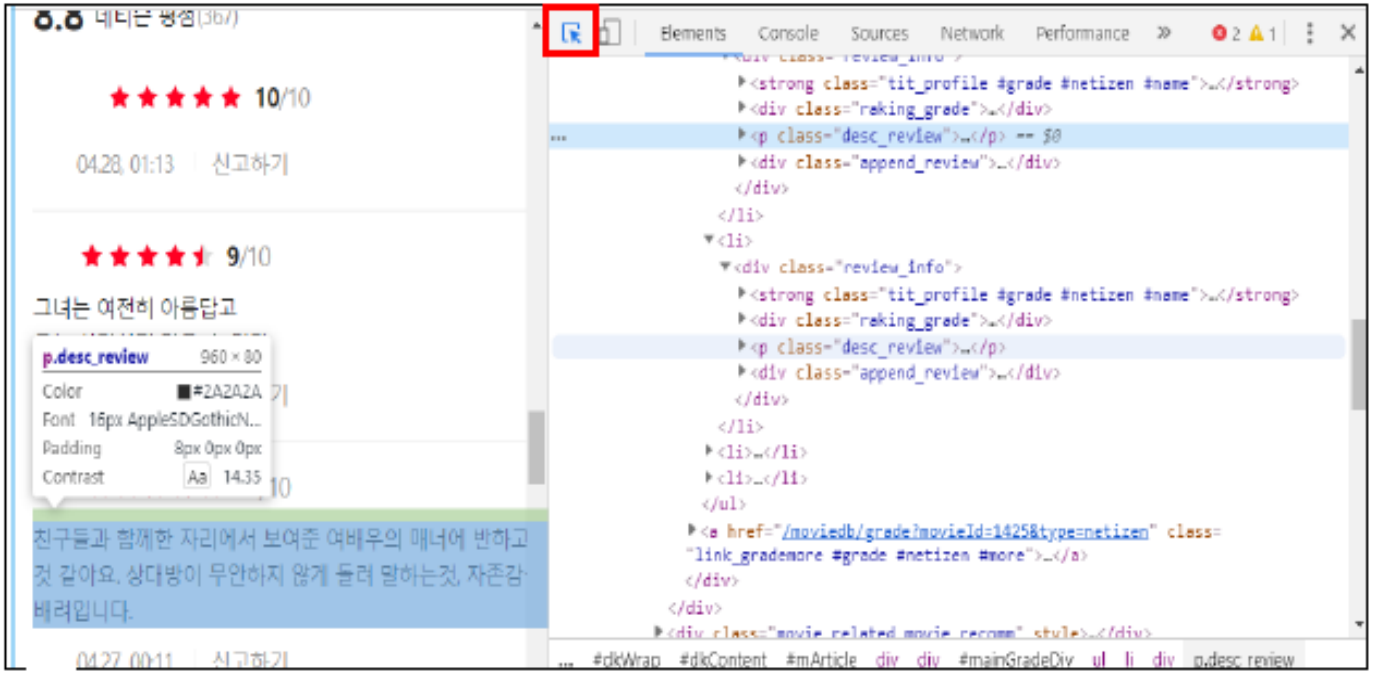
크롬(Chrome) 브라우저의 개발자 도구

- 오른편으로 개발자 도구가 출력되고, ‘Elements’ 탭을 클릭하면 브라우저에서 렌더링되고 있는 웹 페이지의 HTML 소스가 출력 됨



웹 크롤링과 웹 스크래핑

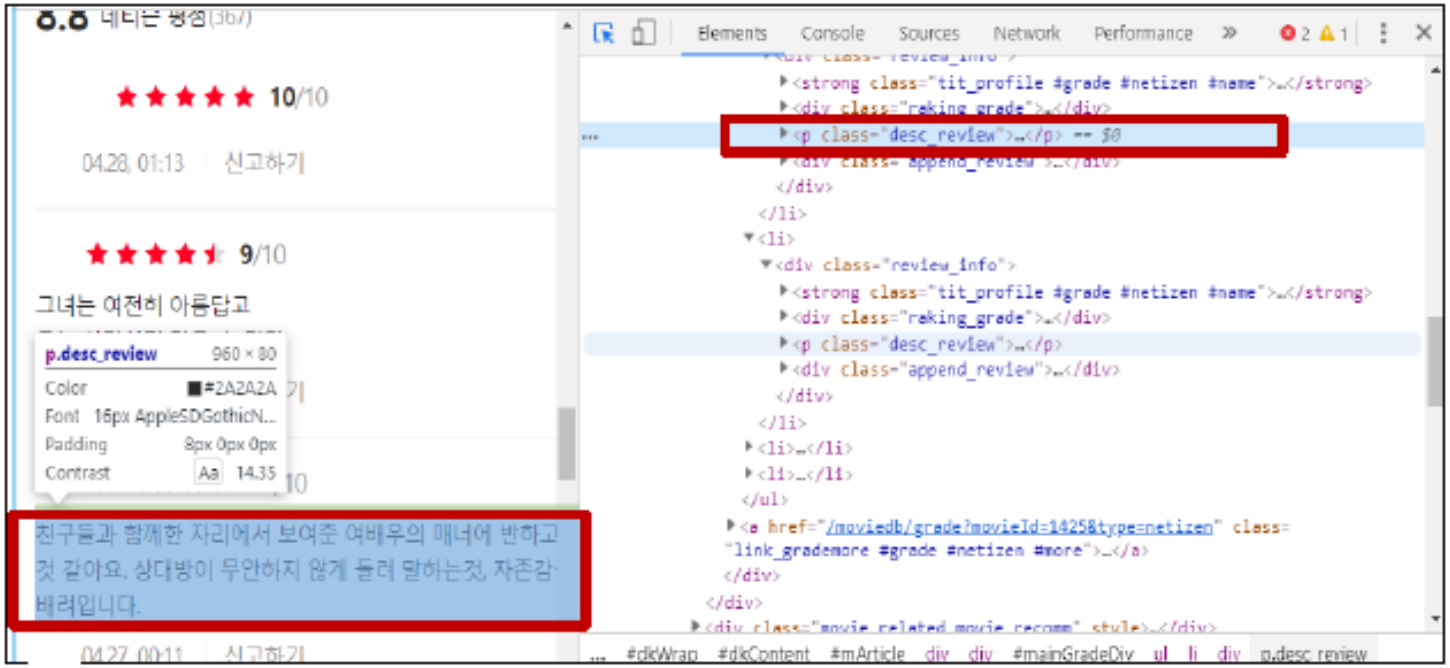
- 크롬(Chrome) 브라우저의 개발자 도구
  - 개발자 도구의 왼쪽 상단 버튼 클릭



웹 크롤링과 웹 스크래핑

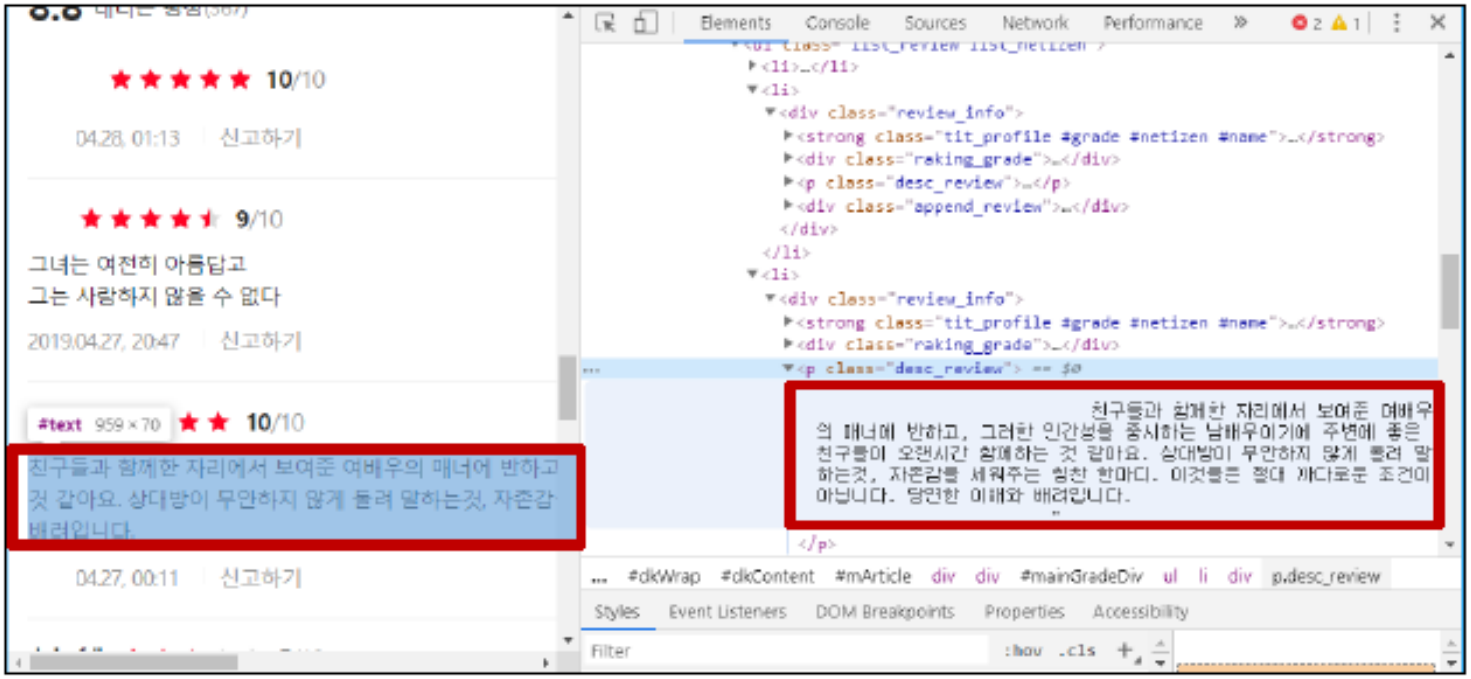
크롬(Chrome) 브라우저의 개발자 도구

- 찾으려는 콘텐츠에 마우스를 올리면, 콘텐츠를 담고 있는 HTML 태그 부분이 개발자 도구의 태그 영역을 표시해주어 찾고자 하는 콘텐츠의 태그를 쉽게 찾을 수 있음



웹 크롤링과 웹 스크래핑

- 크롬(Chrome) 브라우저의 개발자 도구
  - 해당 태그 영역을 클릭하면, 태그의 콘텐츠 영역의 출력 내용이 웹 페이지에 렌더링된 내용과 동일함을 확인할 수 있음



공공 데이터

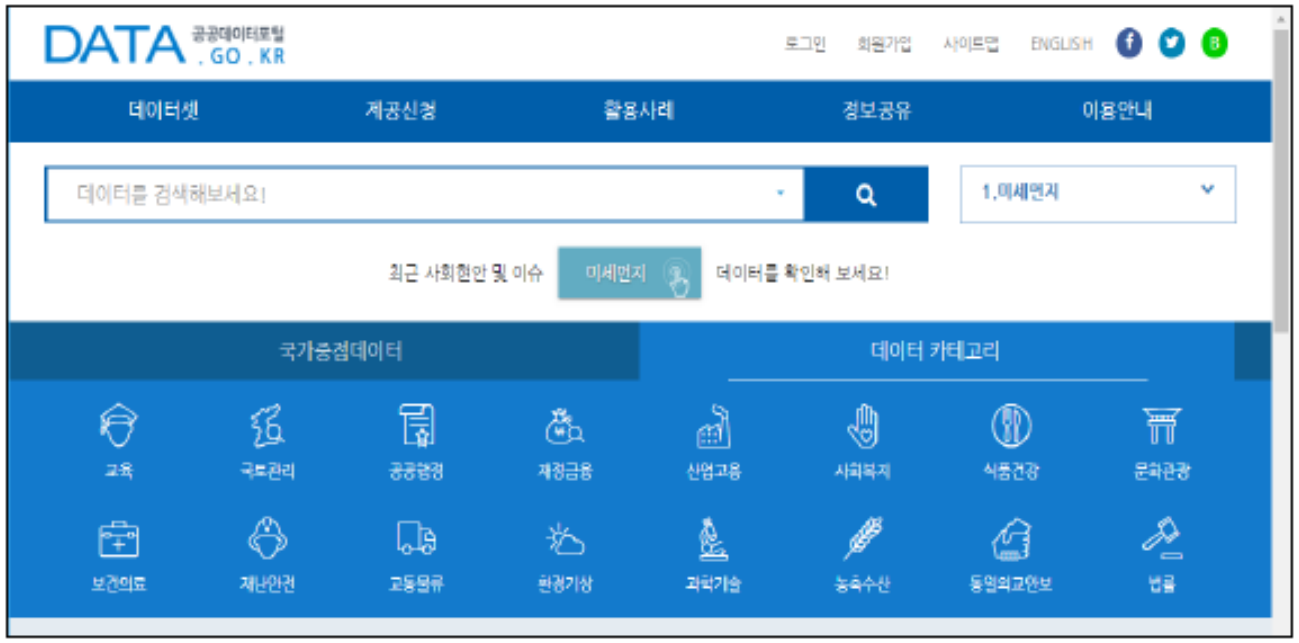
공공데이터	▪ 공공기관이 전자적으로 생성 또는 취득하여 관리하고 있는 모든 데이터베이스(DB), 전자화된 파일
공공데이터 개방	▪ 공공기관이 이용자에게 정보를 재활용할 수 있도록 제공하고, 제공받은 정보를 상업적·비영리적으로 이용할 권한 부여

- 보유하고 있는 공공데이터를 적극적으로 개방하여 국민과 공유함으로써 소통과 협력을 확대하기 위해 공공 데이터 정책 추진
- 2013년 7월 공공 데이터 법을 제정하고 공공 데이터 개방을 10월부터 시행

공공 데이터

공공 데이터 포털(https://www.data.go.kr)

- 공공기관이 생성 또는 취득하여 관리하고 있는 공공 데이터를 한 곳에서 제공하는 통합 창구



공공 데이터

- **서울 열린 데이터 광장(https://data.seoul.go.kr/)**
  - 열린 시정 3.0에 의해 공공 데이터를 민간에 개방하고 소통함으로써 공익성, 업무 효율성, 투명성을 높이고 시민의 자발적 참여로 새로운 서비스와 공공의가치를 창출할 수 있도록 하는 서비스

- **국가통계포털(KOSIS, Korean Statistical Information Service)**
  - 국내, 국제, 북한의 주요 통계를 한곳에 모아 이용자가 원하는 통계를 한번에 찾을 수 있도록 통계청이 제공하는 One-Stop 통계 서비스

- 부산 공공 데이터 포털(https://data.busan.go.kr)
- 부산관광통계 (https://bta.or.kr/mcboard/mn_list.php?mnid=bta&mncd=bta6)

SNS(Social Network Service)

소셜 네트워킹 서비스(Social Networking Service)

- 사용자 간의 자유로운 의사소통과 정보 공유, 인맥 확대 등을 통해
  사회적 관계를 생성하고 강화해주는 온라인 플랫폼
- 최근 스마트폰 이용자의 증가와 무선인터넷 서비스의 확장과 더불어 SNS의 이용자 또한 급증하고 있음

SNS

OPEN API

- 인터넷 이용자가 웹검색 결과 및 사용자 화면 등을 제공 받는데 그치지 않고 직접 응용 프로그램과 서비스를 개발할 수 있도록 공개된 개발자를 위한인터페이스

- 대부분의 SNS 사이트들은 개발자로 등록하고 인증키를 받아 제공되는 API 사용

- **트위터: https://developer.twitter.com/**
- **네이버블로그검색: https://developers.naver.com/docs/search/blog/**
- **네이버뉴스검색: https://developers.naver.com/docs/search/news/**

SNS

RSS(Really Simple Syndication/Rich Site Summary)

- 뉴스나 블로그와 같이 콘텐츠 업데이트가 자주 일어나는 웹사이트에서 업데이트 된정보를 정해진 규격의 XML 형식으로 자동화하여 사용자에게 제공하기 위한 서비스

- RSS가 등장하기 전에는 원하는 정보를 얻기 위해 해당 사이트를 직접 방문 해야 했음
- RSS 관련 프로그램(혹은 서비스)을 이용하여 자동 수집이 가능해졌기 때문에 사용자는 각각의 사이트 방문 없이 최신 정보들만 골라 한자리에서 볼 수 있음

SNS

RSS(Really Simple Syndication/Rich Site Summary)

- 조선일보 RSS(http://rssplus.chosun.com/)
- 인기 뉴스 RSS보기