

ADVANCED TEXT MINING

by FINGEREDMAN (fingeredman@gmail.com)

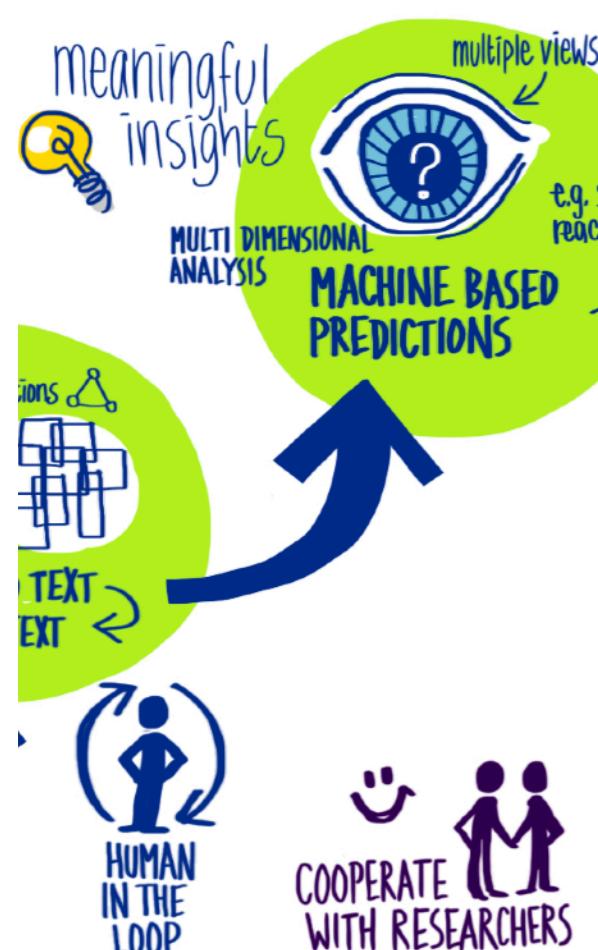
WEEK 03

Web Scrapping

HOW?

아무데서나 찾는 텍스트는
아무 의미없는 정보를 포함한다

텍스트 데이터 수집소스



오프라인 데이터

- 수집방법 : 온/오프라인 설문지, 녹음, 촬영
- 정교한 타겟팅을 통해 원하는 대상 데이터 수집이 가능함
- 사람이 직접 또는 전문업체를 통해 비용을 많이 지불하고 수집해야함
→ 데이터 수집에 시간적, 공간적 제약이 큼

온라인 데이터

- 수집방법 : 웹페이지에 존재하는 모든 데이터
- 대량의 정보를 반복적으로 빠르게 수집할 수 있음
- 수집 대상 웹페이지의 API 호출이나 프로그래밍 언어 활용이 필요함
- 개인정보와 저작권 문제에 취약함

시스템 데이터

- 수집방법 : 게시판, 댓글, 보고서 등 시스템 상에 존재하는 모든 저장된 문서
- 기 수집된 데이터를 바로 활용하여 데이터 수집에 시간적 제약이 없음
- 시스템 관련 부서가 아니거나 관련부서 승인 없이 접근이 어려움
- 내부 데이터를 활용하므로 정보 유출 및 보안관리에 대한 위험성이 큼



비정형 데이터 수집소스

유형별 텍스트 데이터 수집 소스

소스	유형	플랫폼	주요토픽	회원수	사용연령대	성별
디시인사이드	커뮤니티	자체플랫폼	공통		10~30	공통
루리웹	커뮤니티	자체플랫폼	공통		20~30	공통
뽐뿌	커뮤니티	자체플랫폼	공통		10~30	공통
일베저장소	커뮤니티	자체플랫폼	공통		10~40	공통
스레딕	커뮤니티	자체플랫폼	공통		20~30	여성
도탁스	카페	다음	공통	511,049	-	공통
이토랜드	토렌트	자체플랫폼	공통		-	공통
네이트판	커뮤니티	자체플랫폼	고민, 이슈		10~30	공통
오늘의유머	커뮤니티	자체플랫폼	유머		10~30	공통
웃긴대학	커뮤니티	자체플랫폼	유머		10~30	공통
엽기혹은진실	카페	다음	유머	247,754	-	공통
유머나라	카페	다음	유머	114,626	-	공통
와이고수	커뮤니티	자체플랫폼	유머, 스포츠, 게임		10~40	남성
쭉빵카페	카페	다음	연예, 뷰티	1,731,956	20~30	여성
뉴빵카페	카페	다음	연예, 뷰티	1,101,596	20~30	여성
여성시대	카페	다음	연예, 뷰티	729,142	20~30	여성
파우더룸	카페	네이버	뷰티	1,856,696	20~30	여성
인스티즈	커뮤니티	커뮤니티	연예, 오락		10~30	여성
theqoo	커뮤니티	자체플랫폼	연예		10~20	여성
해연갤	커뮤니티	자체플랫폼	해외 연예		20~30	여성
가생이	커뮤니티	자체플랫폼	연예, 한류		20~40	-
베스티즈	커뮤니티	자체플랫폼	연예		20~30	여성
디젤매니아	카페	네이버	패션	882,132	20~30	남성
외방커뮤니티	커뮤니티	자체플랫폼	미용, 패션		20~30	여성

비정형 데이터 수집소스

유형별 텍스트 데이터 수집 소스

소스	유형	플랫폼	주요토픽	회원수	사용연령대	성별
레몬테라스	카페	네이버	육아, 인테리어	3,020,341	30~40	여성
맘스홀릭 베이비	카페	네이버	육아	2,684,457	20~30	여성
개드립	커뮤니티	자체플랫폼	유머, 게임		10~20	공통
인벤	커뮤니티	자체플랫폼	게임		10~20	남성
에펨코리아	커뮤니티	자체플랫폼	축구		10~40	남성
아이러브사커	카페	다음	축구	167,706	10~40	남성
MLB파크	커뮤니티	커뮤니티	야구		20~40	남성
이종격투기	카페	다음	격투기	1,023,757	10~30	남성
클리앙	커뮤니티	자체플랫폼	테크, 통신, 앱		20~40	남성
쿨엔조이	커뮤니티	자체플랫폼	테크, 하드웨어		20~40	남성
Seeko	커뮤니티	자체플랫폼	전자기기		30~40	남성
아사모 - 애플	카페	네이버	애플 아이폰	1,635,061	-	공통
중고나라	카페	네이버	중고거래	16,477,444	10~50	공통
중고카페 그린유즈	카페	네이버	중고거래	2,543,783	-	공통
보배드림	커뮤니티, 쇼핑몰	자체플랫폼	중고거래		30~50	공통
취업뽀개기	카페	다음	취업, 학생	1,399,394	20~30	공통
독취사 - 취업	카페	네이버	취업	2,393,699	20~30	공통
오르비	커뮤니티	자체플랫폼	수험생, 입시		10~20	공통
수만휘	카페	네이버	수험생	2,515,951	10~20	공통
82쿡	커뮤니티, 쇼핑몰		요리		20~50	여성
SLR클럽	커뮤니티	자체플랫폼	사진		30~50	공통
유랑 - 유럽여행	카페	네이버	여행	1,880,443	20~30	공통
네일동 - 일본여행	카페	네이버	여행	1,205,047	20~30	공통
외방커뮤니티	커뮤니티	자체플랫폼	미용, 패션		20~30	여성

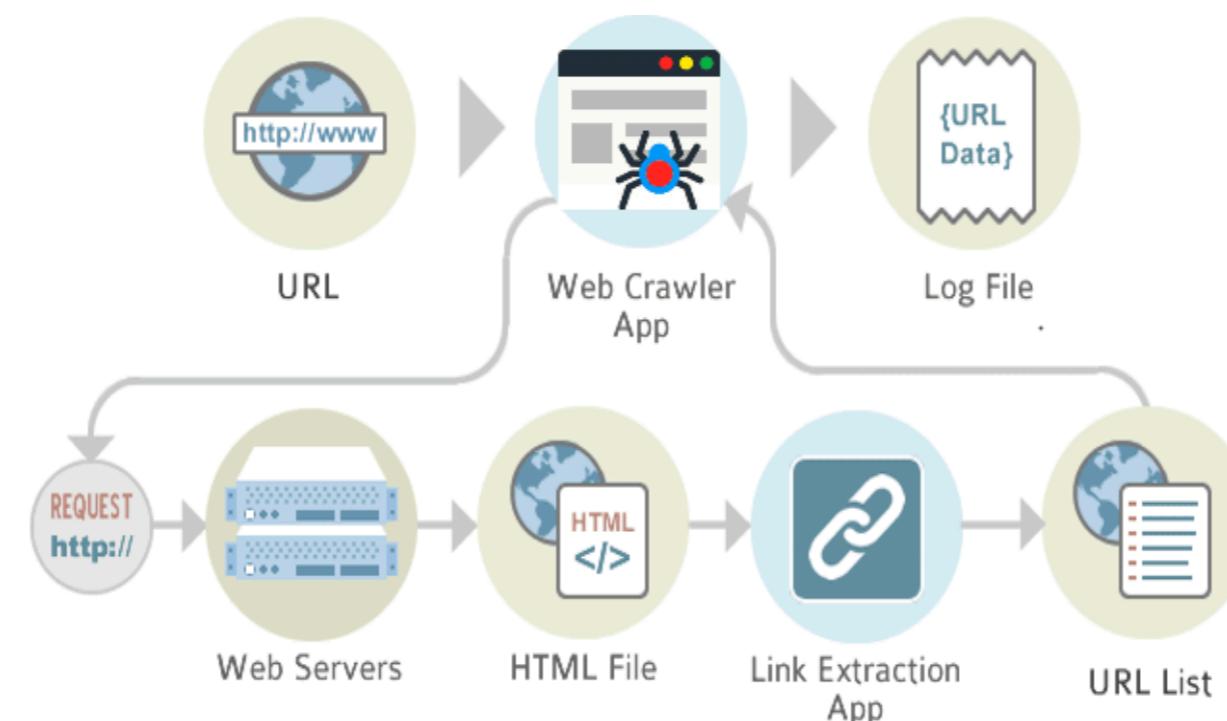
웹 크롤링 & 스크래핑

웹 크롤링(Web Crawling)의 사전적 의미

- 정제되지 않은 웹페이지에서 필요한 데이터를 추출/수집하는 행위
- 즉시 활용 가능한 데이터(파일, 데이터베이스 등)를 제외하고, 웹페이지에 게시된 데이터를 수집하는 기술
- API(Application Programming Interface)를 통해 데이터를 제공하는 경우가 있지만 제약이 많아 웹크롤링을 통한 수집이 요구됨

웹 크롤링 vs 웹 스크래핑, 원론적 의미의 차이점

- **웹 크롤링 (crawling)** : 웹페이지의 하이퍼링크를 돌아다니며 페이지 정보를 추출하고 저장하는 방법
- **웹 스크래핑 (scrapping)** : 웹페이지에서 필요한 정보만 추출하여 저장하는 방법
- 두 가지 모두 웹페이지 정보를 저장하는 방법으로 데이터를 수집하고 활용하기 위한 사전작업에 해당함



마크업 언어 이해

마크업 언어 (Mark-up Language)

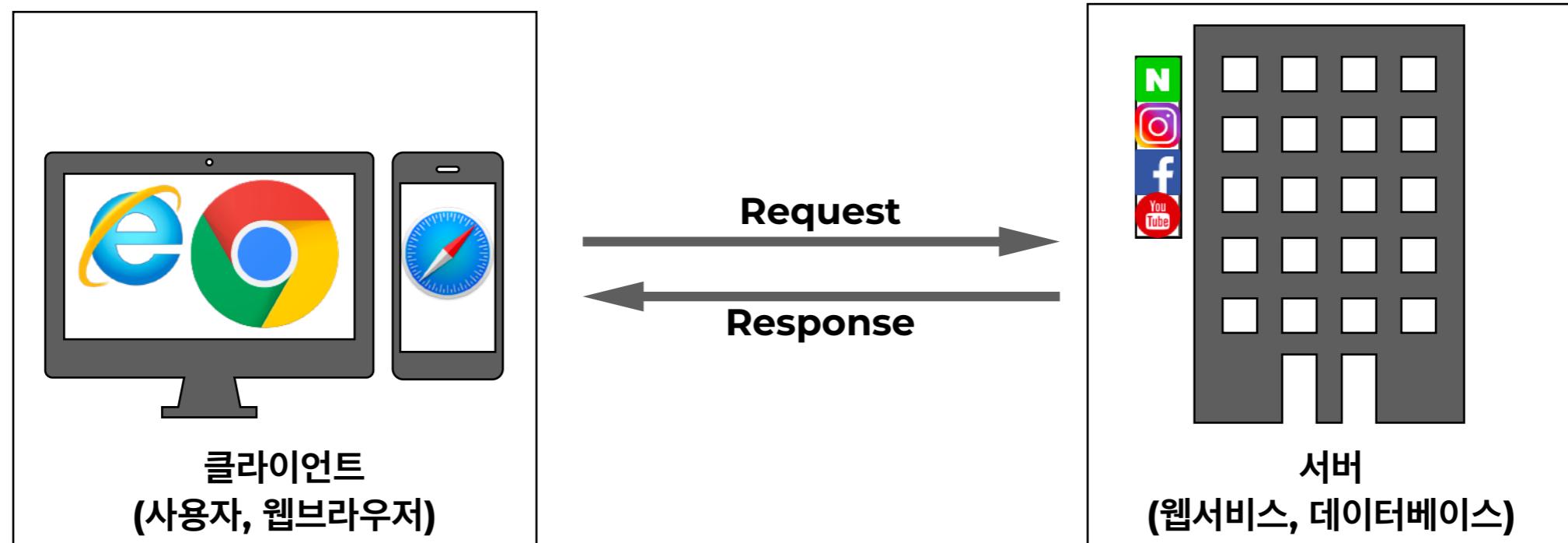
- 태그, 규칙 등을 이용하여 문서나 데이터의 구조를 표현하는 언어 또는 체계
- 태그는 원래 텍스트와는 별도로 원고의 교정부호와 주석을 표현하기 위한 것이었으나 용도가 점차 확장되어 문서의 구조를 표현하는 역할을 하게됨
- 대표적인 마크업 언어 : HTML, JSON, XML, Markdown

XML	JSON	HTML	Markdown
<직원정보>	{"직원정보":	<html>	# H1, 샵이 늘어날수록
<직원리스트>	"직원리스트": [<head>	## H2, 글자크기가 작아집니다.
<직원>	{	<title>공지사항</title>	> 대화블럭을
<이름>서강현</이름>	"이름" : "서강현",	</head>	>> 단계적으로 생성합니다.
<나이>32</나이>	"나이" : 32	<body>	1. 첫 번째 항목
</직원>	,	<p>내일 휴무입니다.</p>	2. 두 번째 항목
<직원>	{	</body>	3. 세 번째 항목
<이름>김민수</이름>	"이름" : "김민수",	</html>	* 1단계
<나이>25</나이>	"나이" : 25		- 2단계
</직원>	}		+ 3단계
</직원리스트>]		+ 4단계 구조를 생성합니다.
</직원정보>	}		

웹 크롤링 & 스크래핑

웹서비스 동작방식: 웹서버 (Web Server)

- 클라이언트 측의 사용자가 웹브라우저(Chrome, Explorer, Safari 등)를 통해 웹서비스에 요청을 보내고 웹서버가 해당 요청에 대해 응답하는 과정의 반복을 통해 동작함



URL <http://www.service.com>

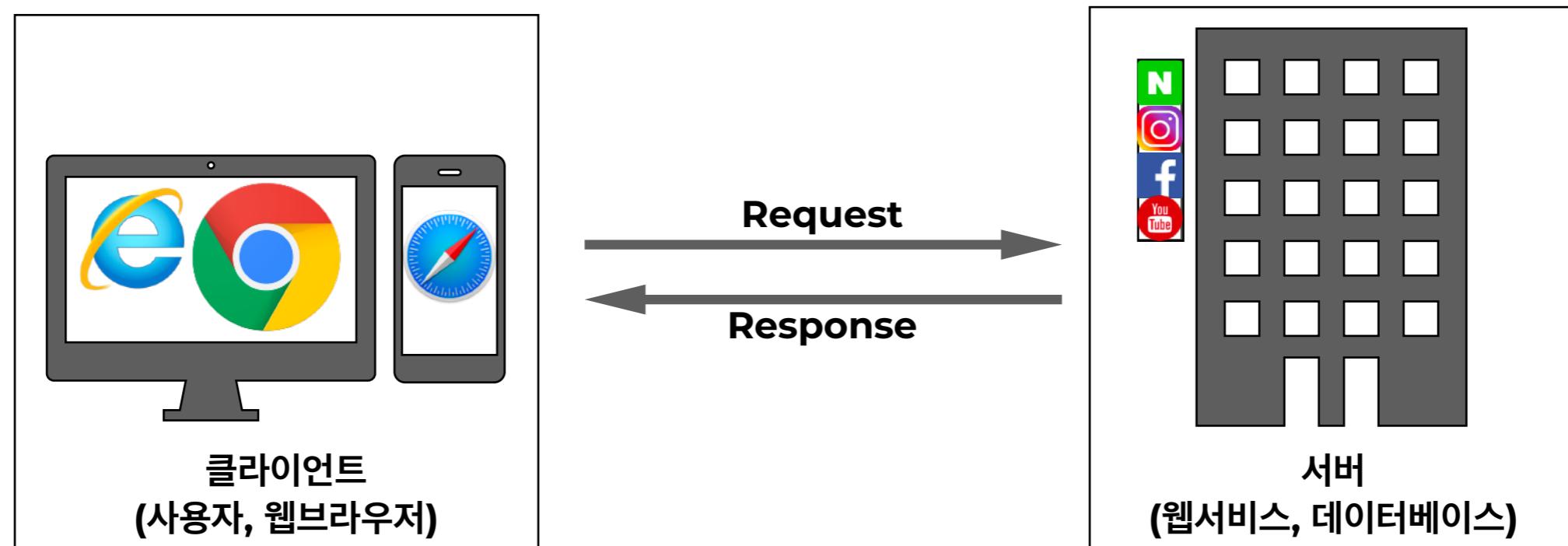
HTML

```
<html>
  <head>
    <title>Hello HTML</title>
  </head>
  <body>
    <p>Hello World!</p>
  </body>
</html>
```

웹 크롤링 & 스크래핑

웹서비스 동작방식: API 서버 (Application Programming Interface Server, API Server)

- 서버에서 제공하는 웹서비스 또는 데이터를 다른 응용 프로그램에서 사용할 수 있도록, 운영체제나 프로그래밍 언어가 기능을 제어할 수 있게 만든 인터페이스



URL <http://api.teanaps.com/nlp/pos>

Data {
 "access_token": "token",
 "sentence": "손흥민은 축구선수 입니다."
}

JSON {
 "code": 200,
 "sentence": "손흥민은 축구선수 입니다.",
 "pos_list": [{"pos": '손흥민', 'pos_tag': 'NNP'},
 {"pos": '은', 'pos_tag': 'JX"},
 ...
]
}

웹 크롤링 & 스크래핑

Python을 활용하는 이유

- 가장 널리 사용되는 프로그래밍 언어로 배우기 쉽고 따라하기 쉬운 장점이 있음
- 풍부한 외부 라이브러리로 프로그래밍이 쉽고 확장성이 좋음
- 데이터 분석과 머신러닝/딥러닝에 가장 많이 사용되는 프로그래밍 언어로, 데이터 수집 뿐만 아니라 수집된 데이터에 대한 전처리, 분석으로 연계하기에 유리함

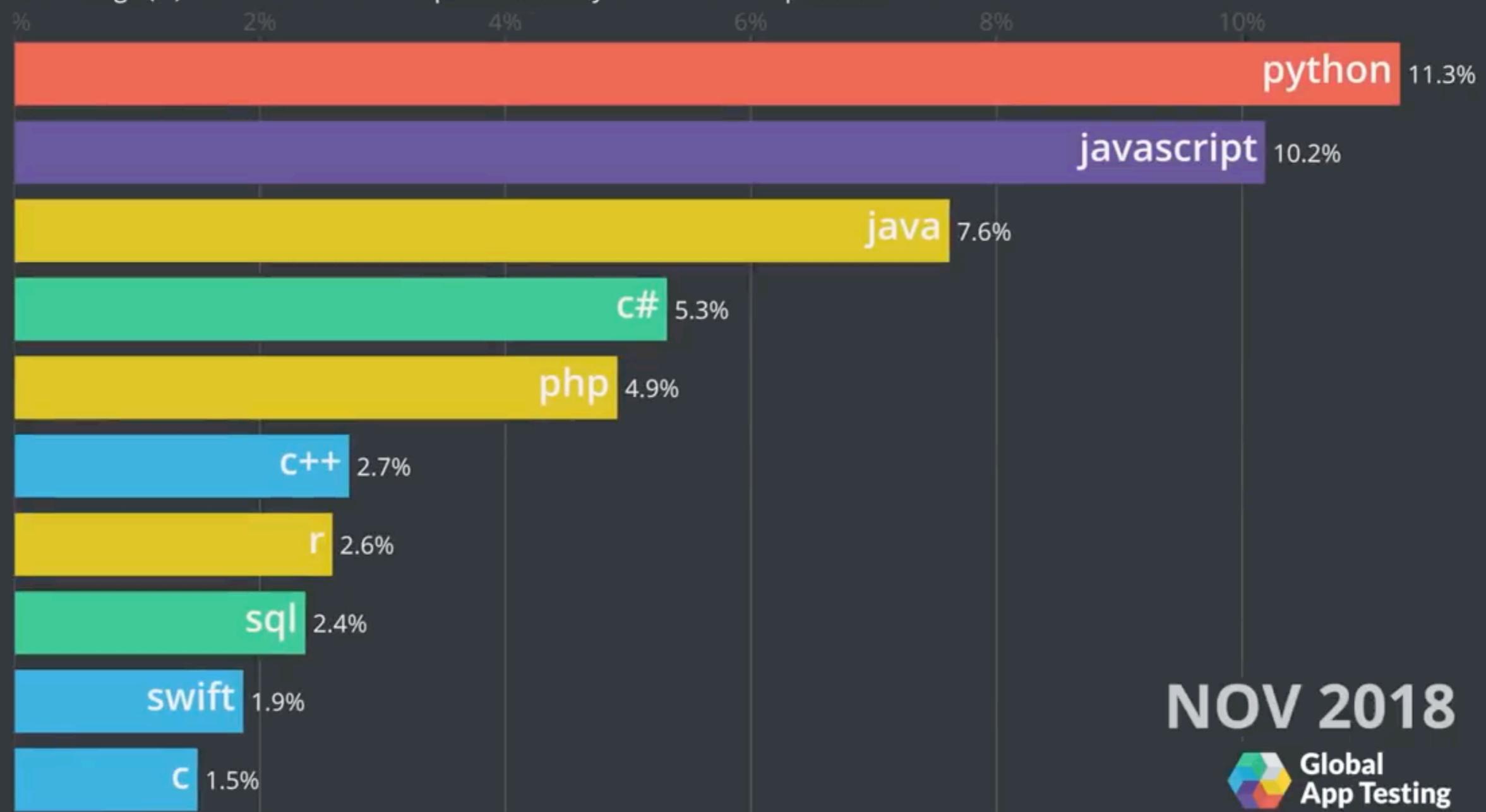
웹 스크래핑 방법 선택

라이브러리	장점	단점
Requests (urllib)	Python에서 동작하는 가장의 작고 빠른 브라우저로, 거의 모든 플랫폼에서 구동 가능 Selenium에 비해 속도가 수백배 빠름	웹서버로부터 초기 HTML만 받아 동적인 처리결과가 반영되지 않음
Selenium	웹브라우저를 Python을 활용해 원격으로 컨트롤 하는 라이브러리 (Chrome, Firefox, IE, PhantomJS 등)	직접 브라우저를 사용하여 동적인 처리를 위한 리소스가 많이 필요함



Most popular programming languages on Stack Overflow.

Percentage (%) of all Stack Overflow questions every month since September 2008.



NOV 2018



웹 크롤링 & 스크래핑

데이터 무단수집과 저작권 침해

- 웹 스크래핑 기술은 원래 검색엔진 등의 인터넷 사이트에서 데이터를 최신 상태로 유지하기 위해 개발됨
- 웹 스크래핑을 활용하여 타사 컨텐츠를 무단 활용하는 것은 불법행위에 해당하며, 과도한 데이터 수집은 대상 웹서비스 운영과 서비스 관리에 안 좋은 영향을 끼침
- 무분별한 웹 스크래핑은 경쟁사 간의 상도덕 문제 또는 개인 양심상의 문제로 확장될 수 있음

채용정보 무단복제 '사람인HR', 잡코리아에 120억 지급

양보다 질 중요한 취업포털 업계… 접근 쉬운 채용공고 속성 악용한 편취사례

이준영 기자 | 승인 2018.02.09 12:31 | 댓글 0

JOBKOREA

saramin

사진=각사

제 및 게재 행위를 하지 않고 공정한 경쟁질서의 확립에 힘쓸 것"이라고 밝혔다.

채용공고 불법 복제 및 게재하는 웹크롤링 행위를 두고 10여 년간 갈등을 빚어온 사람인과 잡코리아가 마침내 합의를 이뤘다.

사람인은 웹크롤링 소송 합의금으로 잡코리아에 120억을 지불했다. 사람인은 이 같은 내용을 공시하고 10일 동안 사람인의 인터넷 웹사이트에 사과문을 공고함으로써 "향후 잡코리아 채용정보 복

댓글부대 의혹 야놀자, 무단 DB 크롤링 의혹 여기어때

숙박 O2O 시장 논란 언제까지

최진홍 기자 | rgdsz@econovill.com | 승인 2017.11.03 16:23:50



모바일 시대가 도래하며 O2O 스타트업의 존재감이 날카로워지고 있지만 잡음 또한 높아지고 있다. 이들은 온라인 경쟁력을 키우면서도 오프라인 거점도 확보, 이를 통한 다양한 파생 서비스에 나선다는 목표도 세워놓고 있다. 그러나 숙박 O2O 업체들이 용인할 수 있는 수준을 넘어설 정도로 구설에 오르고 있는 것은 여간 심각한 문제가 아니다. 최근 나름 적절한 수위를 찾아간다는 평가가 나오고 있지만 배달의민족, 요기요, 배달통 등이 포진한 배달앱 업계도 마찬가지고 다방과 직방 등 부동산 O2O 시장도 사정이 비슷하다. 그 중에서 숙박 O2O 시장을 둘러싼 논란은 상상 이상이다.

웹 크롤링 & 스크래핑

Robots.txt

- 웹 사이트에 배치된 텍스트 파일로, 웹 스크래핑 접근권한에 대해 명시해 놓은 문서
- 원칙적으로 웹 스크래핑은 Robots.txt 파일에서 허용하는 범위 내에서만 가능하며 그 외의 수집에 대한 책임은 모두 본인에게 있음
- 웹 스크래핑이 허용되더라도 대상 웹 사이트 운영에 피해를 주지 않는 선에서 필요한 만큼만 수집해야함
- Robots.txt 파일이 없는 경우 서비스 관리자에 직접 허락을 구한 후 수집해야함

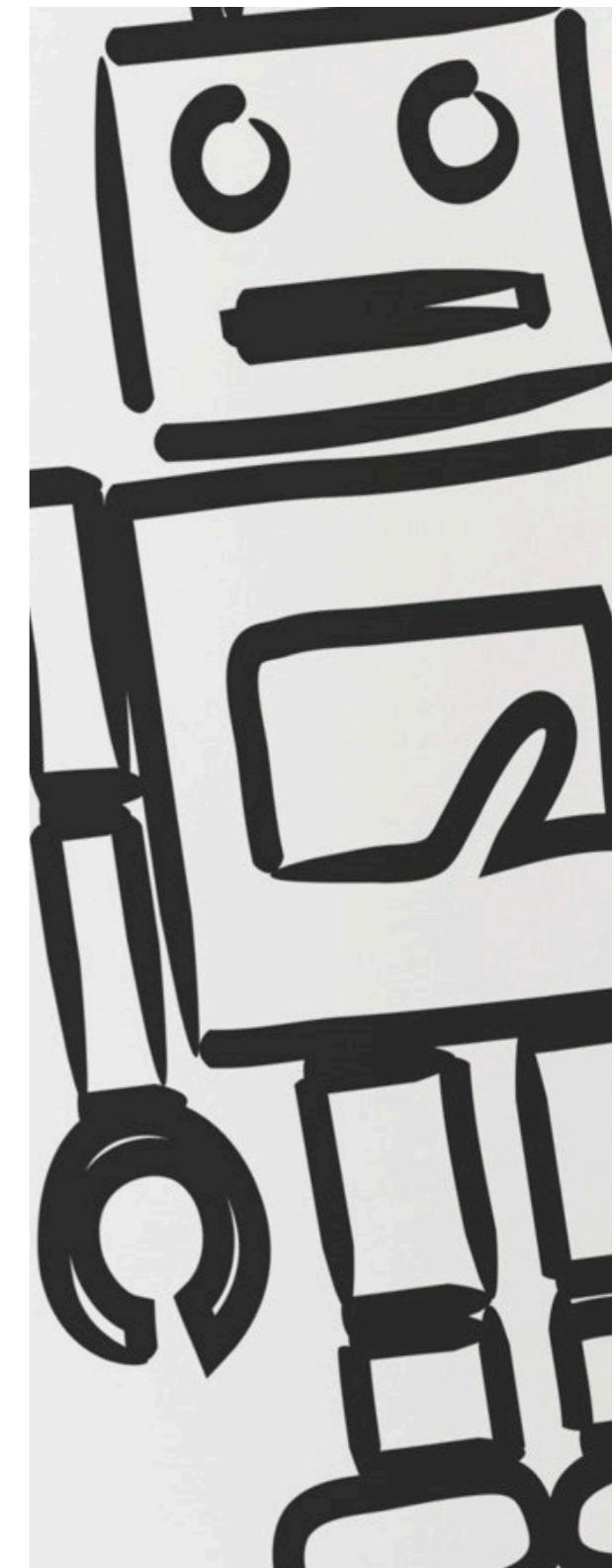


```
User-agent: Bingbot
Allow: /ajax/pagelet/generic.php/PagePostsSectionPagelet
Allow: /safetycheck/

User-agent: Googlebot
Allow: /ajax/pagelet/generic.php/PagePostsSectionPagelet
Allow: /safetycheck/
```



```
User-agent: *
Allow:/service/board/
Disallow:/service/group/
Disallow:/service/board/sold/
Disallow:/service/mypage/
Disallow:/service/message/
Disallow:/service/popup/
Disallow:/service/search/
Disallow:/service/cs/
```



E.O.D

Contact

-  <http://www.teanaps.com>
-  fingeredman@gmail.com