TEXT MINING for PRACTICE

Python을 활용한 비정형 데이터 분석 - WEEK 03

텍스트 데이터 유형 및 수집

연세대학교 | 서중원

텍스트 데이터 수집

텍스트 데이터 수집 유형

- 수집대상 : 웹페이지, SNS, 댓글, 음성, 비디오 등 텍스트 형태로 변환 가능한 모든 데이터
- 저장유형 : Plain Text, PDF, Table, XML, JSON
- 수집방법 : Web Crawling, API 호출, DB Query, Online Survey

유형	수집방법	장/단점
오프라인 데이터	• 온/오프라인 설문지 • 음성 녹음, 비디오 촬영	• 타게팅 대상에 대한 데이터 수집가능 • 사람이 직접 수집해야함 • 수집에 시간적, 공간적 제약이 큼
자체 시스템	• 서비스 또는 사내 데이터베이스 활용 • 사내 게시판, 유저 댓글, 업무 보고서, 내부문서	• 기수집된 데이터를 빠르게 활용 가능 • 소속기관/업체/서비스 내부 관련자가 아니면접근이 어려움 • 정보유출에 대한 위험이 큼
웹크롤링	• 프로그래밍 언어를 활용해 웹페이지에 존재하는 대량의 정보를 반복 수집	• 대량의 정보를 빠르게 수집할 수 있음 • 데이터 수집과 함께 정규화된 데이터셋 구성 가능 • 프로그래밍 언어 활용이 필요함 • 개인정보 문제에 취약함
API 호출	• 프로그래밍 언어를 활용해 서비스에서 정식으로 제공하는 데이터를 수집 • 네이버 API, 카카오 API, Reddit API, News API, SNS (Twitter, Facebook, Instagram)	• 바로 활용할 수 있는 양질의 데이터를 얻을 수 있음 • 수집할 수 있는 소스가 제한적임 • 프로그래밍 언어 활용이 필요함

텍스트 데이터 수집

텍스트 데이터를 제공하는 API

서비스 유형	서비스	제공 유형	비고
SNS	Facebook	• Graph API • Post, Comments	• https://developers.facebook.com 참고
	Instagram	• Media, Comments, Tags	• https://www.instagram.com/developer 참고
	Twitter	• Search API • Streaming API	• https://developer.twitter.com 참고
포털	네이버	• 블로그 API • 뉴스 API • 백과사전 API • 웹문서 API	• https://developers.naver.com 참고 • 본문만 제공
	다음(카카오)	• 웹문서 검색 API • 블로그 검색 API • 카페 검색 API	• https://developers.kakao.com 참고 • 본문만 제공
커뮤니티	Reddit	• Thread, Comment	• https://www.reddit.com/dev/api 참고
뉴스	News API	• Article Summery	• https://newsapi.org 참고
	American Broadcasting Company	• Resources API	• Resources API retrieves content produced by ABC businesses, including national and local news, entertainment videos, and more
	BBC	• Platform API	• Platform API power all the BBC's product areas
	New York Times	• Article Search API • Community API • Movie Reviews API • TimesTags API • Top Stories API	• Headlines, abstracts and links to associated multimedia • Comments by NYTimes.com users • Links to reviews and NYT Critics' Picks, and search movie reviews • Terms that match search queries, and filters by Times dictionaries • Lists of home page articles and associated images

******Source : 전병진, (2019). 소셜미디어 텍스트 마이닝을 위한 통합 애플리케이션 개발 : KoALA. 연세대학교 정보대학원.**

텍스트 데이터 수집

타게팅 유저 텍스트 수집을 위한 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

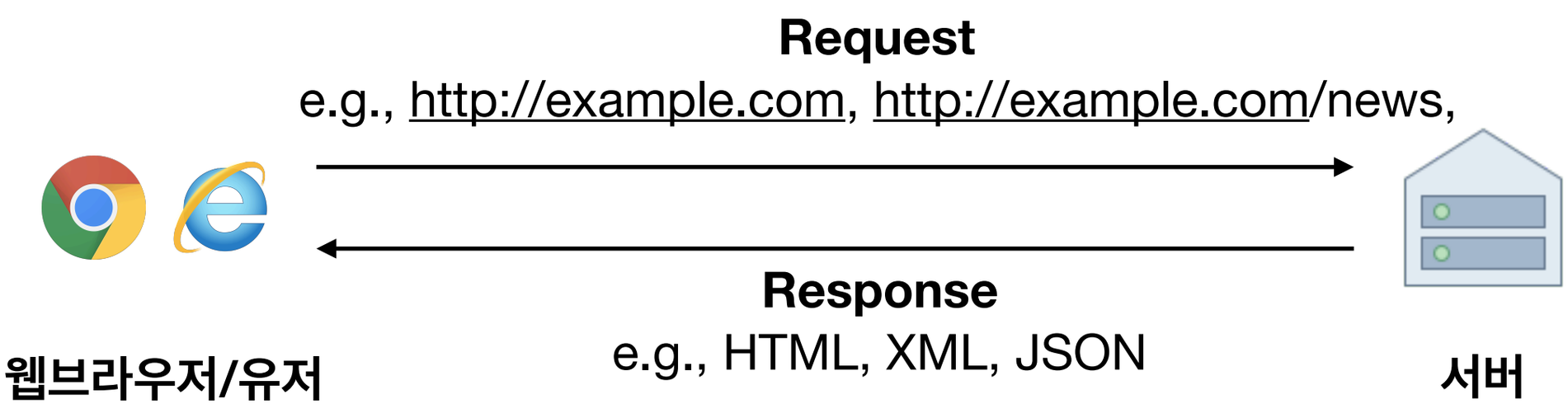
텍스트 데이터 수집

타게팅 유저 텍스트 수집을 위한 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

웹 서비스의 동작 원리

요청 (Request)과 응답 (Response)



Response의 타입

```
<html class="client-js gr__en_wikipedia_org ve-not-available" lang="en" dir="ltr">
▶<head>…</head>
▼<body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-editabl
vector action-view" data-gr-c-s-loaded="true">
    <div id="mw-page-base" class="noprint"></div>
    <div id="mw-head-base" class="noprint"></div>
  ▼<div id="content" class="mw-body" role="main">
      <a id="top"></a>
    ▶<div id="siteNotice" class="mw-body-content">…</div>
      <div class="mw-indicators mw-body-content">
      </div>
    ▶<h1 id="firstHeading" class="firstHeading" lang="en">…</h1>
    ▶<div id="bodyContent" class="mw-body-content">…</div>
    </div>
  ▶<div id="mw-navigation">…</div>
  ▶<div id="footer" role="contentinfo">…</div>
  ▶<script>…</script>
  ▶<script type="application/ld+json">…</script>
  ▶<script>…</script>
  ▶<div class="suggestions" style="display: none; font-size: 13px;">…</div>
    <a accesskey="v" href="https://en.wikipedia.org/wiki/Text_mining?action=edit" cla
  ▶<div id="mwe-popups-svg">…</div>
  </body>
▶<span class="gr__tooltip">…</span>
</html>
```

HTML

```
{
    "currency": "btc",
    "volume": "1696.9926",
    "last": "11314000.0",
    "yesterday_last": "11290000.0",
    "timestamp": "1564408227",
    "yesterday_low": "11136000.0",
    "errorCode": "0",
    "yesterday_volume": "1236.2185",
    "high": "11527000.0",
    "result": "success",
    "yesterday_first": "11286000.0",
    "first": "11290000.0",
    "yesterday_high": "11410000.0",
    "low": "11000000.0"
}
```

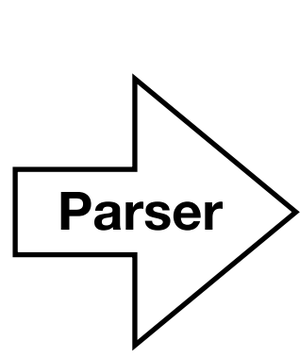
JSON

Data Parsing

Parsing이란?

- 컴퓨터 과학에서 파싱(parsing)은 일련의 문자열을 의미있는 토큰(token)으로 분해하고 이들로 이루어진 파스 트리(parse tree)를 만드는 과정을 말한다.

```
<data>
  <name>서중원</name>
  <major>컴퓨터과학</major>
  <skills>
    <skill>텍스트마이닝</skill>
    <skill>웹프로그래밍</skill>
    <skill>시스템구조</skill>
  </skills>
</data>
```



data
- name: 서중원
- major: 컴퓨터과학
- skills:
  - 텍스트마이닝
  - 웹프로그래밍
  - 시스템구조

Python에서의 데이터 파싱

- JSON
  - 기본 파이썬 패키지를 사용해서 파싱가능
- HTML/XML
  - 파이썬 패키지 Beautiful Soup을 사용해서 파싱

Open API vs 웹 크롤링

Open API 활용

- **장점**: 정제된 데이터(JSON, XML)을 받을 수 있음
- **단점**: 제공하는 데이터만 받을 있음

웹 크롤링

- **장점**: API가 제공하지 않는 웹상에서 볼 수 있는 데이터를 모두 받을 수 있음
- **단점**: 정제되지 않음(HTML), 상업적 이용에 제약이 있음

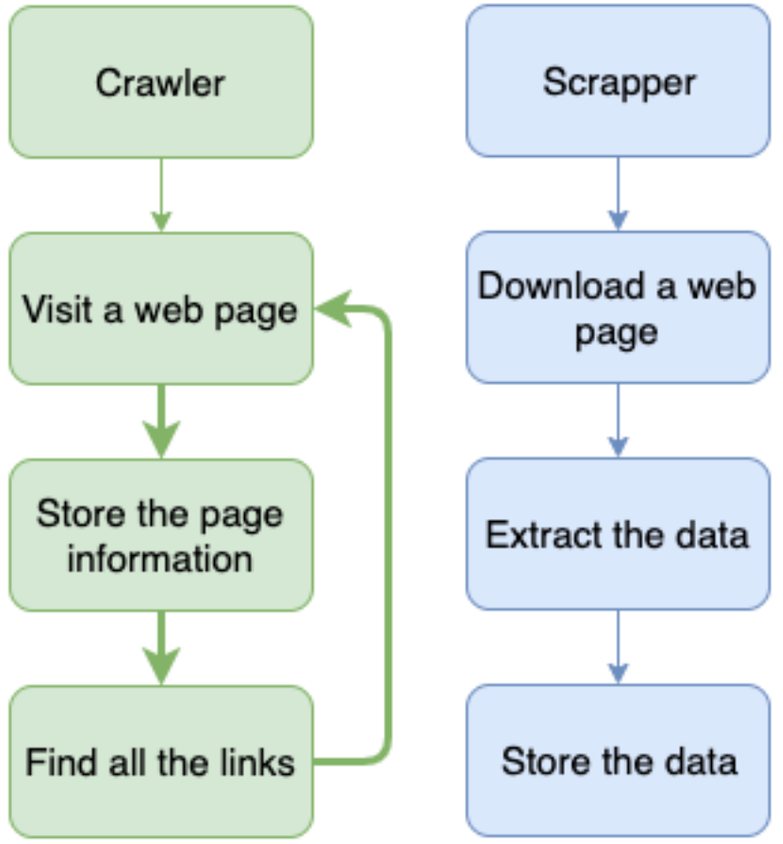
Crawling 이란?

Crawling 이란?

- 정제되지 않은 웹페이지에서 필요한 데이터를 추출하는 행위
- 활용할 수 있는 데이터가 정리되어 올라가 있는 데이터(API, 파일형태)를 제외하고, 웹페이지에 게시된 자료를 가져오는 기술
- API를 통한 데이터 제공이 활성화 되고 있는 추세이나, 국내에서는 소극적인 API 제공으로 웹크롤링을 통한 수집이 반드시 필요함

Crawling과 Scrapping의 차이

- **Crawling**: 웹페이지의 정보를 추출하는 방법(=spider, bot), 단순히 하이퍼링크를 돌아다니며 웹페이지를 다운로드
- **Scrapping**: 웹페이지에서 필요한 정보만을 추출하는 방법, 다운로드한 웹페이지에서 필요한 부분만을 추출하고 저장



크롤링의 법적 문제

데이터 무단수집과 저작권 침해

- 웹크롤링은 원래 검색엔진 등의 인터넷 사이트에서 데이터를 최신 상태로 유지하기 위해 사용
- 웹크롤링을 활용하여 타사 컨텐츠를 무단 활용하는 것은 불법행위에 해당함
- 웹크롤링을 통한 과도한 요청은 대상 서비스 서버 운영과 서비스 관리에 안좋은 영향을 끼침
- 경쟁사간의 상도덕 문제 또는 개인 양심상의 문제

채용정보 무단복제 '사람인HR', 잡코리아에 120억 지급

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

이준영 기자 승인 2018.02.09 12:31 댓글 0

JOBKOREA

saramin

사진= 각 사

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

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.

* Source : 이준영(시장경제), 채용정보 무단복제 '사람인HR', 잡코리아에 120억 지급, 2018.2.9., http://www.meconomynews.com/news/articleView.html?idxno=11088/.
** Source : 최진홍(이코노믹리뷰), 댓글부대 의혹 야놀자, 무단 DB 크롤링 의혹 여기어때, 2017.11.3., http://www.econovill.com/news/articleView.html?idxno=325820/.

크롤링의 법적 문제

Robots.txt

- 웹사이트에 배치된 텍스트 파일로, 크롤링 접근권한에 대해 명시해 놓은 문서
- 웹크롤링은 Robots.txt 파일에서 허용하는 항목에 대해서만 수집 가능하며 그 외의 수집에 대한 책임은 모두 본인에게 있음
- 수집이 허용되 있더라도 대상 서비스 운영에 피해를 주지 않는선 에서 필요한 만큼만 수집
- Robots.txt 파일이 없는 경우 서비스 관리자에 직접 허락을 구한 후 수집

facebook

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

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

CLIEN

```
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/
```

Python 웹 크롤링

Requests를 이용한 크롤링

- URL기반으로 특정 호스트에 요청을 보낸후 응답을 받음
- **장점**: 빠르고 간결, 추가적인 세팅(브라우저 드라이버 설치 등)이 필요하지 않음
- **단점**: Javscript에 의해 추가적으로 요청되는 데이터에 접근하기 어려움

Selenium을 이용한 크롤링

- 웹 브라우저를 원격으로 컨트롤 하면서 데이터에 접근
- **장점**: 사용자가 브라우저를 사용하면서 볼수 있는 모든 데이터에 접근 가능함
- **단점**: 브라우저를 추가적으로 사용하는 만큼 **Requests**에 비해 느리고 추가 세팅이 필요함

E.O.D